

格致方法·定量研究系列　吴晓刚　主编

缺失数据

[美] 保罗·D. 阿利森（Paul D. Allison） 著

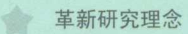
林毓玲　译

革新研究理念

丰富研究工具

最权威、最前沿的定量研究方法指南

29



格致出版社　上海人民出版社



格致方法·定量研究系列



1. 社会统计的数学基础
2. 理解回归假设
3. 虚拟变量回归
4. 多元回归中的交互作用
5. 回归诊断简介
6. 现代稳健回归方法
7. 固定效应回归模型
8. 用面板数据做因果分析
9. 多层次模型
10. 分位数回归模型
11. 空间回归模型
12. 删截、选择性样本及截断数据的回归模型
13. 应用logistic回归分析（第二版）
14. logit与probit：次序模型和多类别模型
15. 定序因变量的logistic回归模型
16. 对数线性模型
17. 流动表分析
18. 关联模型
19.
20. ……应用问题
21.
22.
23. 分析复杂调查数据（第二版）
24. 分析重复调查数据
25. 世代分析（第二版）
26. 纵贯研究（第二版）
27. 多元时间序列模型
28. 潜变量增长曲线模型
29. 缺失数据
30. 社会网络分析（第二版）
31. 广义线性模型导论
32. 基于行动者的模型
33. 基于布尔代数的比较法导论
34. 微分方程：一种建模方法
35. 模糊集合理论在社会科学中的应用
36. 图形代数
37. 项目功能差异

上架建议：社会研究方法

ISBN 978-7-5432-2160-4

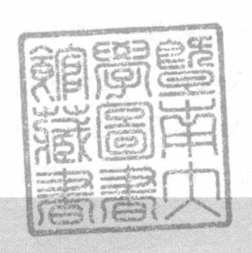


定价：15.00元
易文网：www.ewen.c
格致网：www.hibooks

格致方法 · 定量研究系列　吴晓刚　主编

缺 失 数 据

[美]保罗 · D.阿利森(Paul D. Allison) 著
林毓玲 译



SAGE Publications ,Inc.

格致出版社　上海人民出版社

图书在版编目(CIP)数据

缺失数据/(美)阿利森(Allison, P. D.)著;林毓玲译. —上海:格致出版社:上海人民出版社,2012
(格致方法·定量研究系列)
ISBN 978-7-5432-2160-4

Ⅰ.①缺… Ⅱ.①阿… ②林… Ⅲ.①统计数据-数据处理-研究 Ⅳ.①C819

中国版本图书馆 CIP 数据核字(2012)第 215456 号

责任编辑 高 璇

格致方法·定量研究系列
缺失数据
[美]保罗·D. 阿利森 著
林毓玲 译

出　　版 世纪出版集团 www.ewen.cc　格致出版社 www.hibooks.cn　上海人民出版社
(200001 上海福建中路193号24层)


编辑部热线 021-63914988
市场部热线 021-63914081

发　　行 世纪出版集团发行中心
印　　刷 浙江临安曙光印务有限公司
开　　本 920×1168 毫米 1/32
印　　张 5.25
字　　数 101,000
版　　次 2012 年 10 月第 1 版
印　　次 2012 年 10 月第 1 次印刷
ISBN 978-7-5432-2160-4/C·86
定　　价 15.00 元

出版说明

由香港科技大学社会科学部吴晓刚教授主编的“格致方法·定量研究系列”丛书，精选了世界著名的 SAGE 出版社定量社会科学研究丛书中的 35 种，翻译成中文，集结成八册，于 2011 年出版。这八册书分别是：《线性回归分析基础》、《高级回归分析》、《广义线性模型》、《纵贯数据分析》、《因果关系模型》、《社会科学中的数理基础及应用》、《数据分析方法五种》和《列表数据分析》。这套丛书自出版以来，受到广大读者特别是年轻一代社会科学工作者的欢迎，他们针对丛书的内容和翻译都提出了很多中肯的建议。我们对此表示衷心的感谢。

基于读者的热烈反馈，同时也为了向广大读者提供更多的方便和选择，我们将该丛书以单行本的形式再次出版发行。在此过程中，主编和译者对已出版的书做了必要的修订和校正，还新增加了两个品种。此外，曾东林、许多多、范新光、李忠路协助主编参加了校订。今后我们将继续与 SAGE 出版社合作，陆续推出新的品种。我们希望本丛书单行本的出版能为推动国内社会科学定量研究的教学和研究作出一点贡献。

总 序

往事如烟，光阴如梭。转眼间，出国已然十年有余。1996年赴美留学，最初选择的主攻方向是比较历史社会学，研究的兴趣是中国的制度变迁问题。以我以前在国内所受的学术训练，基本是看不上定量研究的。一方面，我们倾向于研究大问题，不喜欢纠缠于细枝末节。国内一位老师的话给我的印象很深，大致是说：如果你看到一堵墙就要倒了，还用得着纠缠于那堵墙的倾斜角度究竟是几度吗？所以，很多研究都是大而化之，只要说得通即可。另一方面，国内（十年前）的统计教学，总的来说与社会研究中的实际问题是相脱节的。结果是，很多原先对定量研究感兴趣的学生在学完统计之后，依旧无从下手，逐渐失去了对定量研究的兴趣。

我所就读的美国加州大学洛杉矶分校社会学系，在定量研究方面有着系统的博士训练课程。不论研究兴趣是定量还是定性的，所有的研究生第一年的头两个学期必须修两门中级统计课，最后一个学期的系列课程则是简单介绍线性回归以外的其他统计方法，是选修课。希望进一步学习定量研

究方法的可以在第二年修读另外一个三学期的系列课程，其中头两门课叫“调查数据分析”，第三门叫“研究设计”。除此以外，还有如“定类数据分析”、“人口学方法与技术”、“事件史分析”、“多层线性模型”等专门课程供学生选修。该学校的统计系、心理系、教育系、经济系也有一批蜚声国际的学者，提供不同的、更加专业化的课程供学生选修。2001 年完成博士学业之后，我又受安德鲁 · 梅隆基金会资助，在世界定量社会科学研究的重镇密歇根大学从事两年的博士后研究，其间旁听谢宇教授为博士生讲授的统计课程，并参与该校社会研究院（Institute for Social Research）定量社会研究方法项目的一些讨论会，受益良多。

2003 年，我赴港工作，在香港科技大学社会科学部，教授研究生的两门核心定量方法课程。香港科技大学社会科学部自创建以来，非常重视社会科学研究方法论的训练。我开设的第一门课“社会科学里的统计学”（Statistics for Social Science）为所有研究型硕士生和博士生的必修课，而第二门课“社会科学中的定量分析”为博士生的必修课（事实上，大部分硕士生在修完第一门课后都会继续选修第二门课）。我在讲授这两门课的时候，根据社会科学研究生的数理基础比较薄弱的特点，尽量避免复杂的数学公式推导，而用具体的例子，结合语言和图形，帮助学生理解统计的基本概念和模型。课程的重点放在如何应用定量分析模型研究社会实际问题上，即社会研究者主要为定量统计方法的“消费者”而非“生产者”。作为“消费者”，学完这些课程后，我们一方面能够读懂、欣赏和评价别人在同行评议的刊物上发表的定量研究的文章；另一方面，也能在自己的研究中运用这些成熟的

方法论技术。

上述两门课的内容，尽管在线性回归模型的内容上有少量重复，但各有侧重。“社会科学里的统计学”(Statistics for Social Science)从介绍最基本的社会研究方法论和统计学原理开始，到多元线性回归模型结束，内容涵盖了描述性统计的基本方法、统计推论的原理、假设检验、列联表分析、方差和协方差分析、简单线性回归模型、多元线性回归模型，以及线性回归模型的假设和模型诊断。“社会科学中的定量分析”则介绍在经典线性回归模型的假设不成立的情况下的一些模型和方法，将重点放在因变量为定类数据的分析模型上，包括两分类的 logistic 回归模型、多分类 logistic 回归模型、定序 logistic 回归模型、条件 logistic 回归模型、多维列联表的对数线性和对数乘积模型、有关删节数据的模型、纵贯数据的分析模型，包括追踪研究和事件史的分析方法。这些模型在社会科学研究中有着更加广泛的应用。

修读过这些课程的香港科技大学的研究生，一直鼓励和支持我将两门课的讲稿结集出版，并帮助我将原来的英文课程讲稿译成了中文。但是，由于种种原因，这两本书拖了四年多还没有完成。世界著名的出版社 SAGE 的“定量社会科学研究”丛书闻名遐迩，每本书都写得通俗易懂。中山大学马骏教授向格致出版社何元龙社长推荐了这套书，当格致出版社向我提出从这套丛书中精选一批翻译，以飨中文读者时，我非常支持这个想法，因为这从某种程度上弥补了我的教科书未能出版的遗憾。

翻译是一件吃力不讨好的事。不但要有对中英文两种

语言的精准把握能力，还要有对实质内容有较深的理解能力，而这套丛书涵盖的又恰恰是社会科学中技术性非常强的内容，只有语言能力是远远不能胜任的。在短短的一年时间里，我们组织了来自中国内地及港台地区的二十几位研究生参与了这项工程，他们目前大部分是香港科技大学的硕士和博士研究生，受过严格的社会科学统计方法的训练，也有来自美国等地对定量研究感兴趣的博士研究生。他们是：

香港科技大学社会科学部博士研究生蒋勤、李骏、盛智明、叶华、张卓妮、郑冰岛，硕士研究生贺光烨、李兰、林毓玲、肖东亮、辛济云、於嘉、余珊珊，应用社会经济研究中心研究员李俊秀；香港大学教育学院博士研究生洪岩璧；北京大学社会学系博士研究生李丁、赵亮员；中国人民大学人口学系讲师巫锡炜；中国台湾"中央"研究院社会学所助理研究员林宗弘；南京师范大学心理学系副教授陈陈；美国北卡罗来纳大学教堂山分校社会学系博士候选人姜念涛；美国加州大学洛杉矶分校社会学系博士研究生宋曦。

关于每一位译者的学术背景，书中相关部分都有简单的介绍。尽管每本书因本身内容和译者的行文风格有所差异，校对也未免挂一漏万，术语的标准译法方面还有很大的改进空间，但所有的参与者都做了最大的努力，在繁忙的学习和研究之余，在不到一年的时间内，完成了三十五本书、超过百万字的翻译任务。李骏、叶华、张卓妮、贺光烨、宋曦、於嘉、郑冰岛和林宗弘除了承担自己的翻译任务之外，还在初稿校对方面付出了大量的劳动。香港科技大学霍英东南沙研究院的工作人员曾东林，协助我通读了全稿，在此

我也致以诚挚的谢意。有些作者,如香港科技大学黄善国教授、美国约翰·霍普金斯大学郝令昕教授,也参与了审校工作。

我们希望本丛书的出版,能为建设国内社会科学定量研究的扎实学风作出一点贡献。

吴晓刚

于香港九龙清水湾

序

在经验社会科学研究中缺失数据的问题很普遍，大多非实验性研究所报告的统计结果都立足于较小的样本数，有时比初始选择的个案数目还要小。在一些变量上相对多缺失观察值会降低有效的 N。假设有个意见调查，多变量分析中有效样本只有原来的一半，这种情况在现实中并不少见。假设商学院的 Mary Rose 教授在一个消费者态度及行为调查中检验一个 $N=1000$ 受访者的概率样本。她使用一般计算机选项成列删除（也就是任一受访者有缺失任一模型变量即被排除），对支出估计一个合理设定的多元回归模型。结果实际可得的个案降至 $N=499$。这就产生了严重的问题。这 499 个受访者是否仍“代表”了总体？要拒绝零假设，样本是否太小？为了保持样本数，是否应该尝试成对删除？抑或，有其他新的方法值得考虑？这些问题及其他问题都在保罗・阿利森这本杰出的专题著作中讨论到。

“观察值是随机缺失的”，这是根据留下的个案以面对处理数据缺失时的通常论点。但这个假设是隐含的。假若观察值“完全随机缺失”，这表示没有任何变量，不论是因变

量(Y)或自变量(X),其缺失分数都不与该变量自身的值相关。例如,上述的支出变量,对于支出多者其未回答率应不比支出少者的未回答率高。对于其他模型变量,假设在相同的条件下,则499个次样本将可代表一次科学的抽取,允许有效的推论。而且,它允许回归估计值是不偏及一致的。无问题派的研究者可能喜欢这种完全随机缺失的(Missing Completely at Random, MCAR)随机性,但这需要有很强的假设来支持。

较为实际一点的假设为观察值是"随机缺失的"(Missing at Random, MAR)。假设在控制了其他变量后,如果Y的值不能预设缺失分数的位置,则Y变量缺失数据为随机的。所以在上述的举例中,职业地位(X)可能与支出的缺失数据相关,高地位的受访者更可能低报支出。一旦X在右手边,那么Y的观察值将会是随机的。在MAR情况下,如阿利森所言,缺失数据产生机制是可忽略的。虽然他也论及不可忽略的缺失数据机制的困难细节,但他这本专题著作着重于在MAR条件下,以改良估计处理的方法。

如果数据是MAR,则估计的质量很大程度地取决于系统性误差的位置。令人鼓舞的是,当相关缺失数据仅限于自变量时,则成列删除仍能产生不偏的估计值。例如,在例子中,职业地位X缺失数据可能与另一个自变量年龄(Z)相关;例如:没有报告年龄的可能年纪较大且地位较高。在年龄较大与报告支出没有相关的条件下,则没有误差。事实上,正如阿利森巧妙论证的,在一些MAR情况下,标准成列删除选项比传统缺失数据修正方法(成对删除,虚拟变量调整或平均值替换)表现更好。

处理缺失数据问题的新策略占用了本专题论著的大部分篇幅。在缺失数据的条件下回顾最大似然估计,即 ML 估计,他以一个仔细筛选的美国大专院校毕业率的数据为例,解释了插补法的 EM 算法。后几章超越了 ML 方法,解释多重插补方法,并讨论了不可忽略的缺失数据。这本书是最新的处理缺失数据的精心杰作,几乎所有的统计书籍都很少涉及这个主题。保罗·阿利森也睿智地提醒我们,缺失数据最佳的解决方法是"没有任何最佳解决方法"。但如果你也有这个问题且在寻求补救方法,那么就请阅读本书的内容。

迈克尔·刘易斯-贝克

目录

第1章

导论

任何做统计分析的人早晚都会遇到缺失数据的问题。在一个典型的数据组中，对于某些个案的信息是缺失的。例如，在要求个人报告其收入的调查中，通常很大比例的受访者会拒绝回答。彻底拒绝只是造成缺失数据的一个原因。而在自己填写的调查中，人们通常会漏看或忘记回答一些问题。即使是专业的调查员偶尔也会忽略某些问题。有时候受访者说他们只是不知道如何回答或者是没有可用的信息，而有时候某些问题对于一些受访者是不适用的，如让未婚者评价他们的婚姻质量。在长期的研究中，这一波被访问的人在下一波调查前可能会死亡或搬走。当数据从多个行政记录中收集得来时，有些记录可能也会不慎缺失。

因为有这些和其他许多的原因，导致缺失数据在社会和健康科学两者中，成为了一个普遍存在的问题。为什么它会是个问题呢？因为几乎所有标准统计方法都假设每个个案具有可用于分析中的所有变量的信息。确实，大多数的统计教科书没有提到任何有关缺失数据或如何处理缺失数据的信息。

一个众所周知、通常为统计软件默认的、简单的解决方式是：在分析中当某个案的任何变量具缺失数据时，便简单

地将该个案从分析中排除。结果便得到一个没有缺失数据的数据组,可以使用任何传统方法分析。这个策略通常被社会科学称之为成列删除或个案删除,但有时也会被称为完整个案分析。

除了简单以外,成列删除也有一些吸引人的统计特质,这将会在后面讨论。对任何使用过"成列删除"的人而言,它也有一个明显的重大缺点:在很多运用中,成列删除会排除原始样本的一个很大的比例。例如,假使你已收集了一个样本为 1000 人的数据且想要估计一个有 20 个变量的多元回归。每一个变量都有 5%的个案具缺失数据,且每一个变量数据缺失的机会与任何其他变量信息缺失的机会是独立的。那么可预计只有 360 个个案具完整数据,丢弃了其他 640 个"个案"。如果你仅从某网站上下载数据,你可能就不会感觉太糟糕,虽然你也希望能有更多的个案。但如果你已对 1000 人中的每个人都花了 200 美元进行访问,就可能会非常懊悔,因为大概有 13 万美元浪费了(至少对于这个分析是如此)。但在实际操作中,确实有方法可以从这 640 个不完整的个案中抢救回某些东西,因为在这 640 个不完整的个案中,许多个案可能只缺少 20 个变量中的其中一个信息。

许多备选方法被提出来,且其中数个"方法"将在本书中被重新讨论。但遗憾的是,这些方法大多没有价值,且很多未必优于成列删除。虽然这些方法背后的理论已经至少有十年之久,但却仅在过去数年间才可以在计算机上操作。甚至到目前为止,多重插补及最大似然估计仍需要花费大量的时间与精力用于学习方法并根据例行程序执行它们。但如果你想要把事情做对,你通常要付出代价。

最大似然和成列删除两者都有我们希望所能达到的好的统计特性。然而，必须注意的是，这些方法和其他所有的方法一样，其效度基于某些容易被违反的假设。不仅如此，而且没有方法可以检验大部分重要的假设是否被满足。虽然某些解决缺失数据的方法明显优于其他方法，但却没有一个可以被认为确实是好的，也不存在唯一真正解决缺失数据的好的方法。所以在设计和执行研究计划时，必须尽力使缺失信息的发生最小化。因为统计调整没有办法补救草率的研究。

第2章

假 设

研究者通常会试着使那些在某一特别变量上有缺失值的个案与其他有观察值的个案变得没有差异。例如他们通常会提供证据说明报告和不报告其收入的人们在其他各变量上没有显著差异。更普遍地,研究者通常声称或假设他们的数据是随机缺失的,却没有完全理解这代表什么意思。统计学者过去甚至也曾对此概念感到困惑或模棱两可。然而,Rubin (1976)通过严格定义那些关于缺失数据机制不同的合理假设,将这些事物立足于一个坚固的基础上。虽然他的定义相当具技术性,但在此我将试着给出一个通俗的理解。

第1节 | 完全随机缺失的

假设一个特殊变量Y有缺失数据。如果Y数据缺失的概率与Y本身的值或在该数据组中任何其他变量的值都无关的话，那么Y的数据可以说是完全随机缺失的(MCAR)。当这个假设满足所有变量，有完整数据的个人组可被视为原本观察值组中的一个简单随机次样本。要注意，MCAR不考虑Y的"缺失"与其他某个变量X的"缺失"相关之概率。例如，即使拒绝报告其年龄的人们总是拒绝报告他们的收入，但该数据仍然可能是完全随机缺失的。

如果平均而言不报告其收入的人们比那些报告收入者年轻，那么MCAR假设会被违反。很容易就可检验这个暗示，通过将该样本分为两组——报告收入者和未报告收入者，再检验他们的平均年龄的差异。如果实际上有数据呈现者和有数据缺失者两者间在所有观察变量上都没有系统性的差异，则该数据可以算是随机观察的。另一方面，仅由数据通过这个检验并不能说明MCAR假设被满足了，还必须保证某一特殊变量上的缺失与该变量的值没有关系。

虽然MCAR是一个相当强的假设，但有时候它也是合理的，特别是当数据缺失是研究设计的一部分时。当某个特

殊变量测量起来太过昂贵时，这样的设计通常很吸引人。相应的策略便是只针对较大样本中的某个随机次集合，测量这个昂贵的变量，这也意味着对剩余的样本而言，该数据是完全随机缺失的。

第2节 随机缺失的

一个合理性较弱的假设为数据是随机缺失的(MAR)。如果在分析中控制了其他变量后，Y 缺失数据的概率与 Y 值无关，则称 Y 的数据为随机缺失的。为更正式地表达，假设只有两个变量 Y 和 X，X 总是被观察到但 Y 有时会缺失。那么 MAR 指：

$$\Pr(Y \text{ missing} \mid Y, X) = \Pr(Y \text{ missing} \mid X)$$

用文字表示，这个表达式意味着，在同时给定 Y 和 X 时，Y 缺失数据的条件式概率，等于在只单独给定 X 的条件下 Y 缺失数据的概率。例如，如果收入缺失数据的概率取决于婚姻状态，但在每一个婚姻状态类别中，缺失收入的概率与收入无关。一般而言，在控制了其他观察变量后，如果有数据缺失的那些个人相对于那些有数据呈现者，那么对于该变量倾向于较低(或较高)的值，数据不会是随机缺失的。

检验 MAR 条件是否被满足是不可能的，但在直觉上理由应该是很充分的。因为我们不知道缺失数据的值，所以我们无法比较有缺失值者和没有缺失值者，进而看它们是否在该变量上有系统性的差异。

第 3 节 | 可忽略的

如果(a)数据为 MAR,且(b)管制缺失数据过程的参数与要估计的参数无关,则缺失数据的机制是可忽略的(ignorable)。可忽略性基本上指不需要将缺失数据机制模型化为估计过程中的一部分。然而,确实需要使用特别技术以有效地利用数据,因为在实际运用中难以想象条件(b)不被满足的情况,因此作者在本书中将 MAR 和可忽略性视为相等的条件。甚至在极少数条件(b)不被满足的情况下,假设可忽略性的方法仍然运作的一样好,但你可以通过将缺失数据机制模型化从而做得更好。

第 4 节 | 不可忽略的

如果数据不是 MAR，我们则说缺失数据机制是不可忽略的（nonignorable）。在这个例子中，通常缺失数据机制必须被模型化从而得到所关注参数的好的估计值。关于不可忽略缺失数据的一个广泛使用的方法为 Heckman（1976）因变量有选择偏差的两阶段估计回归模型。遗憾的是，对于不可忽略缺失数据的有效估计，需要非常好的关于缺失数据过程本质的基础知识，因为数据没有包含信息告知什么模型是适当的，而且结果会对模型的选择尤其敏感。因为这些原因且因为不可忽略缺失数据的模型要求对每个运用必须相当专业化，因此本书将重点放在可忽略的缺失数据上。在第 7 章，作者简要地分析了一些处理不可忽略的缺失数据的方法。在第 3 章，我们将会看到成列删除有一些非常吸引人的特性，关于某些类型的不可忽略缺失数据这些特性也会非常明显。

第3章

传统的方法

虽然有许多不同的方法被提议用来处理缺失数据，但其中只有一部分得到了广泛好评。但这些被广泛使用的方法中没有任何一个明显优于成列删除。在这部分，我从最简单的方法开始，简短地回顾一些方法。在评估这些方法时，我将特别关注他们在回归分析中的表现（包含 logistic 回归及 Cox 回归），许多评论也适用于其他类型的分析。

第1节 | 成列删除

如前所述,成列删除通过从样本中删除所关注模型中的在任何变量上有缺失值的观察值,并通过运用传统分析完整数据组的方法来实现。成列删除有两个明显的优点:(1)它可以用于任何类型的统计分析,包括从结构方程模型到对数线性分析;(2)并不需要特别的运算方法。根据缺失数据机制,成列删除有一些吸引人的统计特性。确切地说,如果数据为MCAR,则减少的样本将会是原样本的一个随机次样本。这意味着,对于所关注的任何参数,如果估计值对于完整的数据组(没有缺失数据)的估计值是无偏误的,那么对于成列删除的数据组也会是无偏误的。此外,由成列删除的数据组所获得的标准误及检验统计量也如同它们在完整数据组中的一样适当。

当然,因为所利用的信息较少,标准误在成列删除的数据组中通常会比较大。它们也会倾向大于(在本书后面叙述的)由最适当的方法所获得的标准误,但至少你不用担心因为缺失数据而导致推论错误——这是大多数常用方法的一个大问题。

另一方面,如果数据不是MCAR,而只是MAR,那么成列删除可能会产生有偏误的估计值。例如,如果教育缺失数

据的概率取决于职业地位，那么对职业地位进行关于教育的回归将会产生一个有偏误的回归系数估计值。因此，一般而言，成列删除对于违反 MCAR 假设的情况并不是稳健的。但出乎意料的是，成列删除对在回归分析中的自变量间违反 MAR 时是最稳健的。更确切地讲，如果任何因变量缺失数据的概率不取决于自变量的值，则使用成列删除的回归估计值将会是无偏误的(如果所有一般回归模型假设都被满足的话)。[1]

例如，假设我们想估计一个回归模型以预测年储蓄，其中一个自变量为收入，有 40%的数据是缺失的。进一步地假设收入缺失数据的概率取决于收入和教育年数两者，教育年数为模型中另一个自变量。只要缺失收入的概率不取决于储蓄，回归估计值将会是无偏误的(Little, 1992)。

为什么会这样呢？有一个重要的原因。对回归模型的自变量做非比例分层化抽样并不会使系数估计值产生偏误。一个仅与自变量值相关的缺失数据机制在本质上与分层化抽样相同，也就是说，个案以基于其他变量值的概率而被选择进入样本中。这个结论不只适用于线性回归模型，也适用于 logistic 回归、Cox 回归、泊松回归及其他。

事实上，对 logistic 回归而言，甚至在更广泛的条件下，成列删除都能给予有效的推论。如果任何变量缺失数据的概率取决于因变量的值，而不取决于任何其他自变量的值，则使用成列删除的 logistic 回归会产生一致的斜率系数估计值及其标准误(Vach, 1994)。然而，截距估计值将会是有偏误的。只有当任何缺失数据的概率同时取决于因变量和自变量两者时，使用成列删除的 logistic 回归才会是有问题的。[2]

总而言之，成列删除并不是一个差的处理缺失数据的方法。虽然它没有利用所有可得的信息，但至少当数据是MCAR时，它给予了有效的推论。正如我们将看到的，这几乎已经比所有其他普遍处理缺失数据的方法好多了。最大似然和多元插补方法（于后面几节讨论）在许多情况下可能会比成列删除更好，但对于回归分析来说，当违反MAR假设时，成列删除甚至比这些复杂的方法更加稳健。更明确的是，当某个特别的自变量缺失数据的概率取决于该变量（而非因变量）时，成列删除可能会比最大似然和多元插补更好。

对于这些关于回归分析的成列删除的主张有一个重要的提示，即对于样本中的所有个案回归系数都被假设是相同的。如果回归系数在横跨总体的次集合时发生变化，则该样本任何非随机的限制（例如，经过成列删除）都会导致回归系数向其中一个或另一个次集合倾斜。当然，如果在回归参数中察觉到这样的变化，就应该对不同的次样本做不同的回归，或将适当的交互作用包含在模型中（Winship & Radbill，1994）。

第 2 节 | 成对删除

也被称为可得个案分析，成对删除作为一个简单的备选方法可用于许多线性模型，包括线性回归、因子分析及更复杂的结构方程模型。它是广为人知的，例如，一个线性回归可以通过其样本平均数和协方差矩阵，或者通过平均数、标准差及相关矩阵进行估计。成对删除的原理是要通过所有可得的个案来计算这些描述统计的每一个。例如，计算两个变量 X 和 Z 之间的协方差，所有同时具备 X 和 Z 两者数据的个案都会被使用。一旦总和测量值被计算出来了，它们就可用于计算我们所关注的参数，如回归系数。

如何执行这个原则却显得模棱两可。当计算需要每一个变量其平均数的协方差时，只使用这两个变量都有数据的个案来计算平均数吗？还是使用所有变量的数据都可得的个案呢？因为所有的变异都产生具相同特性的估计量，所以毋需考虑诸如此类的问题。一般的结论是，如果数据为 MCAR，成对删除就产生一致的参数估计值（且因此在大样本中接近无偏误）。另一方面，如果数据是 MAR，但不是随机被观察到的，估计值就可能会严重偏误。

如果数据确实是 MCAR，成对删除可能会比成列删除更有效，因为更多信息被利用了。所谓更有效，指的是成对删

除有比成列删除更少的抽样变异(较小的真实标准误)。然而,这并不总是正确的。线性回归模型的分析和模拟研究都指出,当变量间的相关性普遍较低时,成对删除会产生更有效的估计值,然而,当变量间的相关性较高时,成列删除则更好(Glasser, 1964; Haitovsky, 1968; Kim & Curry, 1977)。

成对删除的一个大问题是:由传统软件所产生的标准误和检验统计量估计是偏误的。这个问题的症状是,当你输入一协方差矩阵于回归程序中时,你还必须指明样本数以计算标准误。有些成对删除的程序使用有最多缺失数据的变量的个案数目,而有些程序则用个案的最小数目以计算每一个协方差。然而,没有一个数目是令人满意的。原则上,有可能可以得到标准误的一致估计值,但公式很复杂,且目前在任何商业统计软件都无法执行。[3]

成对删除第二个偶尔会发生的问题是,在小样本中,建构的协方差或相关矩阵可能不是"正定的",这也暗示着回归运算根本无法实行。由于存在这些困难以及其对 MCAR 的偏离相对敏感,因此成对删除通常不被建议为成列删除的备选方法。

第3节 | 虚拟变量调整

这里还有另一个针对回归分析中缺失预测值的非常简单且直觉上很吸引人的方法(Cohen & Cohen, 1985)。假设某变量 X 有一些缺失数据,X 为回归分析中数个自变量的其中一个。我们建立一个虚拟变量 D,如果 X 数据缺失等于1,如果没有缺失等于0。我们也建立一个变量 X^*,使得

$$X^* = \begin{cases} X, \text{当数据没有缺失时} \\ c, \text{当数据缺失时} \end{cases}$$

其中 c 可以是任何常数。我们回归因变量 Y 于 X^*、D 及其他在预设模型中的所有变量。这个技术被称为虚拟变量调整或缺失指标方法,该方法可以很容易被延伸至超过一个自变量具缺失值的数据中。

虚拟变量调整方法明显的好处在于它使用了所有可用的关于缺失数据的信息。将用 c 值代替缺失数据视为插补并不恰当,因为 X^* 的系数不会因为 c 值的不同选择而改变。而且,此模型唯一一个取决于 c 值的选择的面向为缺失值指标 D 的系数。为便于解释,一个简单可选的 c 值是非缺失个案的 X 的平均数。这样 D 的系数可以被解释为,在控制了模型中的其他变量的情况下,X 具缺失数据的个体其 Y 的预

测值减去具 X 平均数的个体其 Y 的预测值。X^* 的系数可被视为在有 X 数据的次群体中 X 的效应的估计值。

遗憾的是，如 Jones(1996)所证明的，这个方法通常会产生有偏误的系数估计值。[4]一个简单的模拟就能说明这个问题。从具有三个变量的正态分布中抽样产生 1 万个个案，这三个变量为 X、Y 和 Z。回归 Y 于 X 和 Z，对于每个自变量，其真实的系数为 1.0。不出意外，利用整个具有上万个个案的样本而得到的最小二乘回归系数（如表 3.1 第 1 栏所示）相当接近真实值。

表 3.1　使用三种模型之仿真数据回归

系　数	完整数据	成列删除	虚拟变量调整
X	0.98	0.96	1.28
Z	1.01	1.03	0.87
D			0.02

其次，再随机地使 Z 值有 1/2 的概率缺失。因为缺失数据的概率与其他任何变量不相关，所以该数据为 MCAR。表 3.1 第 2 栏显示成列删除产生的估计值相当接近那些没有数据缺失的结果。另一方面，虚拟变量调整方法的系数则明显有偏误——X 系数高而 Z 系数太低。

另外有一个密切相关的方法被提议用于回归分析中的类别自变量。诸如此类的变量可通过建立一组虚拟变量来有代表性地处理这个问题，除了参照组外，每一个类别都有一个变量。这个方法是简单地建立一个额外的类别及一个额外的虚拟变量，以表示在该类别变量中具缺失数据的个体。然而，我们还有一个直觉上吸引人但即使当数据为 MCAR 时仍有偏误的方法（Jones，1996；Vach & Blettneer，1991）。

第4节 | 插补

许多处理缺失数据的方法都归在插补方法的大标题之下，其基本原理是要以某些合理的猜测插补或替代缺失值，然后再接着按没有缺失数据的情况进行分析。当然，有许多不同的插补缺失值的方法。最简单的可能是边际平均数插补：对给定某个变量的每一个缺失值，都用有数据个体的该变量的平均数代替。众所周知，这个方法会产生有偏误的方差及协方差的估计值（Haitovsky，1968），因此通常应该避免使用。

一个比较好的方法是利用多元回归的方法使用其他变量的信息，这个方法通常被称为条件式平均数插补。假设我们要估计一个有着数个自变量的多元回归模型。其中一个自变量 X，部分个案有缺失数据。对于那些有完整数据的个案，我们回归 X 于所有其他自变量上。使用相应的估计方程，我们会得到预测值用于具 X 缺失数据的个案。这些值用来代替缺失数据，并且可以接着按没有缺失数据的情况进行分析。

当超过一个自变量有缺失数据而且一般主题有许多变异时，这个方法就会变得比较复杂。一般而言，如果插补是全然根据其他自变量（而非因变量）且数据为 MCAR 时，最

小二乘系数是一致的，暗示在大样本中几乎无偏误（Gourieroux & Monfort, 1981）。然而，它们并不是完全有效的。可通过加权最小二乘（Beale & Little, 1975）或广义最小二乘（Gourieroux & Monfort, 1981）获得改良的估计量。

遗憾的是，所有这些插补方法都面临一个根本的问题：即按照完整数据的情况分析插补数据会低估标准误、高估检验统计量。传统分析分法没有单纯地对这个实际情况（即插补过程涉及缺失值的不确定性）进行调整。[5]后面几节会介绍一个解决这些问题的插补方法。

第 5 节 | 总结

所有从具缺失数据的个案中挽救信息的一般方法都会明显地使事情变得更糟：它们会引进重大的偏误，使分析对MCAR的偏离更加敏感，并产生不正确的标准误（通常太低）。由于存在这些缺点，成列删除因而看起来不算太糟。但还有更好的可用的方法。在下一章中，将会介绍可用于许多一般模型化目标的最大似然方法。在第 5 章和第 6 章，也会介绍几乎可用于任何设定的多重插补方法。如果数据是MAR，那么这两种方法都有不一般的特性。原则上，这些方法可用于不可忽略的缺失数据中，但需要一个有关数据缺失过程的正确的模型——这通常难以得到。

第4章

最大似然

最大似然是统计估计非常普遍的方法，广泛用于处理许多困难的估计问题。大多数读者可能熟悉 ML 作为估计 logistic 回归模型的偏好方法。当误差项假设为正态分布时，普通最小二乘线性回归也是一个 ML 方法。结果 ML 特别有利于处理缺失数据问题。在这一章，我们首先回顾 ML 估计值的一些普遍特性。然后，将介绍缺失数据机制为可忽略的假设下，ML 估计的基本原则。这些原则会用一个简单的列联表来说明。此章其余的部分会给出更复杂的例子，目的是要根据多变量正态分布估计一个线性模型。

第1节｜回顾最大似然估计法

ML估计的基本原则是，选取那些若取值是真实的，就可最大化观察到事实上已被观察到的概率的值，来作为估计值。为了达到这个目的，我们首先需要一个以数据和未知参数两者的函数来表达数据的概率的公式。当观察值为独立(一般的假设)时，该样本的总体似然(概率)就是所有个别观察值的似然的乘积。

假设我们要估计一个参数θ。如果$f(y \mid \theta)$为给定某个θ值时，观察到Y单一值的概率，则一个有n个观察值的样本的似然值为：

$$L(\theta) = \prod_{i=1}^{\pi} f(y_i \mid \theta)$$

其中$\prod$是重复乘法运算的符号。当然，我们仍需要确切指明$f(y \mid \theta)$是什么。例如，假设Y是一个二分类变量，编码为1或0，且θ为$Y=1$的概率，则：

$$L(\theta) = \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i}$$

一旦我们有了称为似然函数的$L(\theta)$，就有许多方法确定尽可能使似然值最大化的θ值。

ML估计量有许多令人满意的特性，已知在相当广泛的

条件下，它们是一致的、渐近有效且渐近正态的（Agresti & Finlay, 1997）。一致性意味着估计值在大样本中接近无偏误。有效性意味着真实的标准误至少和任何其他相应的估计量的标准误一样小。

渐近部分意指这项陈述只是接近真实的，且样本量越大则越接近。最后，渐近正态性意指在重复抽样中，估计值接近正态分布（同样，接近程度会随样本数增加而增加）。这说明使用一个正态表来建构置信区间或计算 p 值是恰当的。

第2节｜有缺失数据的ML

当某些观察值具缺失数据时怎么办？当缺失数据机制是可忽略的时(因此为MAR)，我们可以简单地通过加总所有缺失数据可能值的一般似然来获得似然。例如，假设我们要对一个具n个独立观察值的样本收集两个变量X和Y的数据。对于前m个观察值，我们观察到X和Y两者，但对于剩下的$n-m$个观察值，我们仅能测量到Y。对于有完整数据的单一观察值，我们用$f(x, y \mid \theta)$表示其似然值，其中θ为一组支配X和Y分布的未知参数。假设X是离散的，一个具X缺失数据的个案其似然就是Y的“边际”分布：

$$g(y \mid \theta) = \sum_{x} f(x, y \mid \theta)$$

当X为连续时，加总以积分代替。整个样本的似然便为：

$$L(\theta) = \prod_{i=1}^{m} f(x_i, y_i \mid \theta) \prod_{i=m+1}^{n} g(y_i \mid \theta)$$

问题就变成了寻找尽可能使这个似然值最大化的θ值。许多方法可用于解决这个最适化问题，后面我将介绍其中的一些方法。

当缺失数据的模式为单调时，ML特别简单。在一个单调的形态中，可以用一个顺序安排变量，以使样本观察值的

排列情况为：如果数据对于某特别变量有缺失的话，那么必然地对于排在这个变量之后的其他变量数据也是缺失的。

举一个包含有四个变量 X_1、X_2、X_3 和 X_4 的例子。X_1 没有缺失数据，X_2 100%的个案有缺失数据。缺失了 X_2 的个案在 X_3 和 X_4 这两个变量中数据也有缺失。额外地有20%的个案 X_3 和 X_4 有缺失数据，但 X_2 没有缺失数据。一个单调的形态通常出现于面板研究中，其中，个人在不同时点退出且不再回到研究/数据中。

如果只有一个变量有缺失数据，该形态必然是单调的。考虑一个只有 X 具缺失数据的包含两个变量的例子。$f(x, y)$的联合分布可被写成 $h(x \mid y)g(y)$，其中 $g(y)$ 为 Y 的边际分布（前面已定义过）且 $h(x \mid y)$ 为给定 Y 时 X 的条件式分布。这样我们可把似然值重写为：

$$L(\lambda, \phi) = \prod_{i=1}^{m} h(x_i \mid y_i ; \lambda) \prod_{i=1}^{n} g(y_i \mid \phi)$$

这个表达式与前一个表达式在两个重要方面有所不同。首先，第二个乘积是针对所有观察值的，而不只是那些有 X 缺失数据的观察值的。其次，参数已被分成两部分：λ 表示给定 Y 时 X 的条件式分布，而 ϕ 则表示 Y 的边际分布。这些改变意味着我们可以分别最大化似然值的这两个部分。因此，如果 X 和 Y 有一个二变量正态分布，我们就可以计算整个样本的 Y 的平均数和方差。再者，对于有 X 数据的个案，我们可以回归于 Y 上。得到的参数估计值可以结合起来以产生我们可能会关注的任何其他参数的 ML 估计值，如回归系数。

第3节 | 列联表数据

这些 ML 估计的特征可以用列联表数据非常具体地说明。假设一个简单随机样本 200 人，我们要测量两个取值可能为 1 或 2 的二分类变量 X 和 Y。有 150 个个案，我们观察到 X 和 Y，并将获得的数据列于下面的列联表中：

	$Y=1$	$Y=2$
$X=1$	52	21
$X=2$	34	43

对于另 50 个个案，X 为缺失的，我们只观察到 Y：确切地说我们有 19 个个案 $Y=1$，31 个个案 $Y=2$。在总体中，X 和 Y 的关系被叙述为：

	$Y=1$	$Y=2$
$X=1$	p_{11}	p_{12}
$X=2$	p_{21}	p_{22}

其中，p_{ij} 为 $X=i$ 且 $Y=j$ 的概率，如果我们只有具完整数据的 150 个观察值，似然值为：

$$L=(p_{11})^{52}\,(p_{12})^{21}\,(p_{21})^{34}\,(p_{22})^{43}$$

受到限制，此四个概率必须加总为 1。此四个概率的 ML 估

计值是每一个单元格中的简单比率，也就是：

$$\hat{p}_{ij} = \frac{n_{ij}}{n}$$

其中 n_{ij} 为落入单元格(i, j)中的个案数目。因此我们得到：

$$\hat{p}_{11} = 0.346$$
$$\hat{p}_{21} = 0.227$$
$$\hat{p}_{12} = 0.140$$
$$\hat{p}_{22} = 0.287$$

但这不可行，因为我们还有额外的仅有 Y 的观察值，需要把它们整合到似然值中。假设缺失数据为可忽略的，有 $Y = 1$ 的个案其似然值就是 $p_{11} + p_{21}$，即 $Y = 1$ 的边际概率。同样地，对于 $Y = 2$ 的个案，似然值为 $p_{12} + p_{22}$。因此，整个样本的似然值为：

$$L = (p_{11})^{52}(p_{12})^{21}(p_{21})^{34}(p_{22})^{43}(p_{11} + p_{21})^{19}(p_{12} + p_{22})^{31}$$

我们如何找到使这个表达式最大化的 p_{ij} 值？对大部分关于缺失数据问题的 ML 运用而言，这些估计值并没有明确的解法。然而，迭代的方法是必要的。但在这个例子中，形态必定是单调的（因为只有一个变量有缺失数据），因此，我们可以分别估计给定 Y 时 X 的条件式分布及 Y 的边际分布。然后，我们结合所有结果以得到四个单元格的概率。对于 2×2 的表，ML 估计量的一般形式为：

$$\hat{p}_{ij} = \hat{p}(X = i \mid Y = j)\,\hat{p}(Y = j)$$

右边的条件式概率仅利用完整数据的个案来估计。它们可以通过一个普遍的方法——即将 2×2 表中单元格频数除以

栏总数——而获得。Y 的边际概率估计值可由加总栏频数及缺失值的个案的 Y 频数，再除以样本数而获得。因此，我们得到：

$$\hat{p}_{11} = \left(\frac{52}{86}\right)\left(\frac{86+19}{200}\right) = 0.3174$$

$$\hat{p}_{21} = \left(\frac{34}{86}\right)\left(\frac{86+19}{200}\right) = 0.2076$$

$$\hat{p}_{12} = \left(\frac{21}{64}\right)\left(\frac{64+31}{200}\right) = 0.1559$$

$$\hat{p}_{22} = \left(\frac{43}{64}\right)\left(\frac{64+31}{200}\right) = 0.3591$$

当然，这些估计值与只用具完整讯息个案所得到的估计值不相同。另一方面，一个普遍使用的二分类变量相关性的测量的交叉相乘比，不管是从 ML 估计值计算而来还是仅根据完整个案而得的估计值，都会是相同的。简言之，有 X 缺失数据的观察值没有为我们提供关于交叉相乘比的额外信息。

这个例子用于说明具缺失数据的 ML 估计的一些普遍特征。然而，很少读者会在应用 ML 估计法时，像以上所讲述的用手计算他们的特别运用。我们需要一个可以处理较多数据形式以及缺失数据型态的通用的软件。虽然对于分析列联表的 ML 估计计算并不困难（Fuch，1982；Schafer，1997)，事实上也没有商业软件可以处理这个任务。但在网络上可获取免费的软件：

(1) Jeroen K. Vermunt 的 Windows 版 ℓ_{EM} 程序，用以估计当某些数据缺失时各种类型的类别数据模型。

(http://www.kub.nl/fac-ulteiten/fsw/organisatie/departmenten/mto/software2.htlm)

(2) Joseph Schafer 的 CAT 程序，用以估计缺失数据的分层对数线性模型，但目前只在 S-PLUS 软件包中作为一个程序库可供利用。

(http://www.stat.psu.edu/~jls)

(3) David Duffy 的 LOGLIN 程序，用以估计缺失数据的多种对数线性模型。

(http://www2.qimr.edu.au/davidD)

第4节 | 具正态分布数据的线性模型

在数据来自多元变量的正态分布的假设下，ML 可以用来估计许多线性模型。可能的模型包括普通线性模型、因子分析、联立方程和具潜变量的结构方程。虽然多元变量正态性的假设很强，但对于没有缺失数据的变量是全然无害的。此外，甚至当有些变量具非正态分布的缺失数据时（如虚拟变量），多元变量正态性假设下的 ML 估计值通常有好的特性，尤其是当数据为 MCAR 时。[6]

有许多方法可以用来对具可忽略缺失数据机制的多元变量正态数据做 ML 估计。当缺失数据服从一单调形态时，可用先前叙述过的方法将似然因子化运用到用传统软件估计的条件式及边际分布中（Marini，Olsen & Rubin 1979）。然而，这个方法因可能的运用而受到较多限制，而且不容易得到好的标准误及检验统计量的估计值。

一般缺失数据模式可用一个称为期望最大化的算法来处理（Dempster，Laird & Rubin，1977），它可产生平均数、标准差（或同样地可产生平均数和协方差矩阵）和相关性的 ML 估计值。这些总和统计量接着可被输入到标准线性模型化软件中以得到所关注参数的一致估计值。EM 方法的优点有：(1)它很容易使用；(2)很多商业的或免费的软件都可以

执行。同时也有两个缺点:(1)由线性模型化所报告的软件标准误和检验统计量并不正确;(2)对于过度识别模型,估计值不是全然有效的("过度识别模型"指那些在协方差矩阵上隐含限制的模型)。

一个比较好的方法是直接最大化所假设的线性模型的多元变量正态似然值。直接的 ML(有时也称为原始最大似然)提供具正确标准误的有效估计值,但需要有难以掌握的专业的软件。在本章接下来的部分,我们会具体讲述如何使用 EM 算法和直接 ML。

第 5 节 | EM 算法

当有些数据缺失的时候,EM 算法是取得 ML 估计量的一个非常普遍的方法(Dempster et al. , 1977; McKachlan & Krishnan, 1997)。它之所以被称为 EM,是因为它包含两个步骤:一个期望步骤,一个是最大化步骤。这两个步骤在一个迭代的过程中多次重复,最终收敛到 ML 估计值。

在此我并不解释一般设定下 EM 算法的两个步骤,而是着重于它对于多变量正态分布的运用。这里步骤 E 即期望步骤实质上变成缺失值的回归插补。假设数据组有四个变量,X_1 到 X_4,且每个变量都含有不具特殊形态的缺失值。我们从选择未知参数的起始值,也就是平均数和协方差矩阵开始。这些起始值可以通过样本平均数和协方差的标准公式获得,不管是用成列删除或成对删除。根据参数的起始值,我们可以得到回归任何一个 X 于其他三个 X 上的系数。例如,假设有一些个案 X_1 和 X_2 都有数据,但 X_3 和 X_4 没有数据。我们用协方差矩阵的起始值以得到回归 X_3 于 X_1 和 X_2 及回归 X_4 于 X_1 和 X_2 的结果。我们接着用这些回归系数,根据 X_1 和 X_2 的观察值,产生 X_3 和 X_4 的插补值。对于只有一个变量数据缺失的个案,我们根据所有其他三个变量进行回归插补。

在所有缺失数据都插补完之后，步骤 M 即最大化步骤包括使用插补数据和没有缺失的数据，来计算新的平均数和协方差矩阵的值。对于平均数，我们只要使用一般公式即可。对于方差和协方差，必须使用修正过的公式给每个涉及缺失数据的公式项。确切地讲，必须根据插补过程中的回归方程，对应残差方差和残差协方差从而增加公式项。例如，假设对观察值 i，X_3 为使用 X_1 和 X_2 所插补。那么，只要 $(x_{i3})^2$ 被用于传统的方差公式中，我们就用 $(x_{i3})^2 + s_{3\cdot21}^2$ 代替，其中，$s_{3\cdot21}^2$ 为从回归 X_3 于 X_1 和 X_2 而来的残差方差。加入残差项可以矫正通常在更传统插补方案中产生的对于方差的低估。假设对于观察值 i，X_4 也是缺失的。那么当计算 X_3 和 X_4 的协方差时，只要 $x_{i3}x_{i4}$ 被用于传统协方差公式中，我们就用 $x_{i3}x_{i4} + s_{34\cdot21}$ 代替。最后一项为 X_3 和 X_4 在控制 X_1 和 X_2 后的残差协方差。

一旦我们得到新的平均数和协方差矩阵的估计值，我们就重新开始步骤 E。也就是说，我们用新的估计值来产生对于缺失值的新的回归插补。我们一直循环步骤 E 和步骤 M 直到估计量收敛，即从一个迭代到另一个迭代之间结果已经几乎不变。

注意，EM 运算法避免了传统回归插补的一个难题——决定使用哪些变量作为自变量，并处理不同数据缺失模式有不同的自变量的情况。因为 EM 总是从完整协方差开始，因此它可能得到对于任何组预测量的回归估计值，不论在某个特别的缺失数据形态中可能存在多么少的个案。因此，EM 总是使用所有可得的变量作为预测量以插补缺失数据。

第6节 | EM实例

我们使用来自1994年美国新闻和世界报道对美国最好大专院校的指南中关于美国1302所大专院校的数据。我们考虑下列变量：

GRADRAT　高年级毕业生对四年前就读人数的比率(100%)

CSAT SAT　语言和数学两个科目的平均分数

LENROLL　新生入学人数的自然对数

PRIVATE　1 = 私立;0 = 公立

STUFAC　师生比率(100%)

RMBRD　每年食宿总支出(千美元)[7]

ACT　平均ACT分数

我们的目标是要估计一个线性回归模型，GRADRAT是因变量，其他五个是自变量。ACT不会在回归模型中，它被纳入EM估计是因为它与CSAT(CAST为有大量缺失数据的变量)高度相关，因此可以让我们得到较好的缺失插补值。

表 4.1　利用可得个案而得的大专学院数据的描述性统计

变　量	非缺失个案	平均数	标准差
GRADRAT	1204	60.41	18.89
CSAT	779	967.98	123.58
LENROLL	1297	6.17	1.00
PRIVATE	1302	0.64	0.48
STUFAC	1300	14.89	5.19
RMBRD	783	4.15	1.17
ACT	714	22.12	2.58

表 4.1 列出了每一个变量没有缺失数据的个案数目，以及这些有数据的个案的平均数与标准差。只有 PRIVATE 这个变量的数据是完整的。自变量 GRADRAT 有 8%的大专院校缺失数据。CSAT 和 RMBRD 都缺失 40%左右，而 ACT 则缺失 45%。对除了 ACT 外的所有变量使用成列删除会产生一个只有 455 个个案的样本，明显令人无法接受。然而，为了比较，我们还是把成列删除回归估计值呈现于表 4.2。

表 4.2　使用成列删除预测 GRADRAT 的回归

变　量	系　数	标准误	t 统计量	p 值
截距	−35.028	7.685	−4.56	0.0001
CSAT	0.067	0.006	10.47	0.0001
LENROLL	2.417	0.959	2.52	0.0121
PRIVATE	13.588	1.946	6.98	0.0001
STUFAC	−0.123	0.132	−0.93	0.3513
RMBRD	2.162	0.714	3.03	0.0026

其次，我们使用 EM 运算法来获得平均数、标准差和相关性的估计值。在主要的商业套装软件中，EM 对数法可得于 BMDP、SPSS、SYSTAT 及 SAS。然而，使用 SPSS 和

SYSTAT,要存盘输入其他线性模型化例行程序,这非常麻烦。对于大专院校数据,我们使用 SAS 的 MI 程序,结果显示于表 4.3 和表 4.4。如使用其他 EM 软件,这个程序自动实现前面叙述过的所有步骤。

表 4.3 来自 EM 算法的平均数和标准差

变 量	平均数	标准差
GRADRAT	59.86	18.86
CAST	957.88	121.43
LENROLL	6.17	0.997
PRIVATE	0.64	0.48
STUFAC	14.86	5.18
RMBRD	4.07	1.15
ACT	22.22	2.71

表 4.4 来自 EM 算法的相关性

	GRADRAT	CAST	LENROLL	PRIVATE	STUFAC	RMBRD	ACT
GRADRAT	1.000						
CAST	0.591	1.000					
LENROLL	−0.027	0.192	1.000				
PRIVATE	0.398	0.161	−0.619	1.000			
STUFAC	−0.318	−0.315	0.267	−0.368	1.000		
RMBRD	0.478	0.479	−0.016	0.340	−0.282	1.000	
ACT	0.598	0.908	0.174	0.224	−0.293	0.484	1.000

比较表 4.3 和表 4.1 的平均数可见,不出意料,最大的差异出现在所有有着最多缺失数据的变量中:GRADRAT、CSAT、RMBRD 及 ACT。然而,即使对这些变量而言,成列删除和 EM 结果之间的差异也都没有超过 2%。

表 4.5 显示了使用 EM 统计量得到的回归估计值。虽

然系数与使用成列删除的表 4.2 并没有明显差异，但所得的标准误却低很多，从而导致较高的 t 统计量和较低的 p 值。遗憾的是，在这个例子中虽然系数为真实 ML 估计值，但标准误确实太低，因为它们假设所有个案都有完整数据。为了得到正确的标准误估计值，我们将使用随后叙述的直接 ML 方法。[8]

表 4.5 根据 EM 算法预测 GRADRAT 的回归

变 量	系 数	标准误	t 统计量	p 值
截距	−32.395	4.355	−7.44	0.0001
CSAT	0.067	0.004	17.15	0.0001
LENROLL	2.083	0.539	3.86	0.0001
PRIVATE	12.914	1.147	11.26	0.0001
STUFAC	−0.181	0.084	−2.16	0.0312
RMBRD	2.404	0.400	6.01	0.0001

第 7 节 | 直接 ML

如我们之前看过的，大多数 EM 运算法的软件产生平均数及不受限制的相关性（或协方差）矩阵的估计值。当这些总和统计量被输入其他线性模型化程序时，产生的标准误估计值将会有偏误，而且通常是被低估的。为了做得更好，我们需要直接把所关注模型的似然值最大化。可以使用任一估计包含潜变量的结构方程模型（SEMs）的软件包来完成这个任务。

当只有少量的缺失数据时，可以使用处理多组的任一 SEM 程序来估计线性模型（Allison，1987；Muthen，Kaplan & Hollis，1987），其中包含 LISTREL 和 EQS。对于具更普遍形态的缺失数据，目前有四个程序可以执行线性模型的直接 ML 估计：

Amos　一个 SEM 模型化的商业软件，现在可以作为一个独立的套装或 SPSS 的一个模块来使用。信息参见 http://smallwaters. com。

Mplus　一个独立的商业软件。信息参见 http://statmodel. com。

LINCS　Gauss 的一个商业模块。信息参见 http://www. aptech. com ≠ 3party. com。

Mx　一个免费软件，可从 http://views. vcu. edu/mx 下载。

在进行实例讲解前，让我们先回顾一下基本的理论。令 $f(x \mid \mu, \Sigma)$ 为一个被观察向量 x、平均数向量 μ 及协方差矩阵 Σ 的多变量正态密度。如果我们对于一个来自这个多变量正态分布、有着 $i = 1, \cdots, n$ 个观察值的样本有完整数据，那么似然函数则为：

$$L(\mu, \Sigma) = \prod_i f(x_i \mid \mu, \Sigma)$$

现在假设我们没有完整数据。如果对于个案 i 有一些变量缺失数据，我们令 x_i 为较小的向量，直接将缺失的元素从 x 中去除。令 μ_i 为 μ 的次向量，以排除在 x_i 缺失的对应的元素，且令 Σ_i 为 Σ 的次矩阵，以删除对应于 x 缺失值的列和栏。我们的似然函数则变为：

$$L(\mu, \Sigma) = \prod_i f(x_i \mid \mu_i, \Sigma_i)$$

虽然这个函数看起来非常简单，但却比具完整数据的似然函数更加难以处理。然而，这个似然函数可以由传统的 ML 估计方法来最大化。同时，我们可以取此似然函数的对数，对未知参数偏微分，并令结果等于 0。得到的方程可以使用数学算法来求解，如使用产生标准误如同一个相乘组合的 Newton-Raphson 方法。它也可能将一结构加于 μ 和 Σ 上，从而让它们成为一个相对应于某个假设线性模型的、有较小组参数的函数，例如，因子模型组

$$\Sigma = \Lambda\Phi\Lambda' + \Psi$$

其中 Λ 为因子载荷矩阵，Φ 为潜在因子的协方差矩阵，Ψ 为误差成分的协方差矩阵。估计过程可以产生这些参数的 ML 估计值和标准误估计值。

第 8 节 | 直接 ML 实例

在此用同时具有使用者图解界面及文字界面两者的 Amos 3.6 来估计大专院校的回归模型。图解界面允许使用者用变量中的箭头来指明方程式。因为作者不能做实时示范，故而将相同的文字命令显示于图 4.1 中。数据在一个叫

```
$Sample size = 1302
$missing = -9
$input variables
   gradrat
   csat
   lenroll
   private
   stufac
   rmbrd
   act
$rawdata
$include = c:\college.dat
$mstructure
   csat
   lenroll
   private
   stufac
   rmbrd
   act
$structure
   gradrat = () + csat + lenroll + private + stufac
            + rmbrd + (1)error
   act<>error
```

图 4.1　Amos 预测 GRADRAT 的回归模型指令

COLLEGE. DAT 的自由格式文字文件中，缺失数据标示为－9。$mtructure 告诉 Amos 要估计指定变量的平均数，这是估计有缺失数据模型的一个必要的部分。$structure 指令指明要估计的方程。紧接在等号后的括号代表要估计一个截距。方程最后面的(1)error 告诉 Amos 要包含一个系数为 1.0 的误差项。最后一行，act<>error，考虑 ACT 和误差项之间的相关性，这是有可能的，因为 ACT 对于 GRADRAT 没有直接效应。Amos 自动考虑 ACT 和回归方程中的其他自变量相关。

结果显示于表 4.6。与表 4.5 中两步骤 EM 估计值的比较显示，两者得到的系数相同，但 Amos 标准误显著较大，这正是我们所期望的。但相对于我们在表 4.2 由成列删除所得到的仍然相当小。

表 4.6　以 Amo 使用直接 ML 预测 GRADRAT 的回归

变　量	系　数	标准误	t 统计量	p 值
截距	－32.395	4.863	－6.661	0.000000
CSAT	0.067	0.005	13.949	0.000000
LENROLL	2.083	0.595	3.499	0.000467
PRIVATE	12.914	1.277	10.114	0.000000
STUFAC	－0.181	0.092	－1.968	0.049068
RMBRD	2.404	0.548	4.386	0.000012

第9节 | 结论

对于随机缺失的数据而言，最大似然可以说是一个有效且实用的方法。在这种情况下，对于大样本，ML 估计值应该是最合适的。对于可用如 LISREAL 程序所估计的一般的结构方程模型之线性模型，ML 估计值很容易从许多广泛可用的软件包中获得。类别数据的对数线性模型 ML 估计也有可用的软件，但在这个设定中执行比较间接。ML 方法的一个限制条件为，它需要一个包含所有缺失变量的联合概率的模型。对这个目的而言，多变量正态模型通常是方便的，但对于许多其他的运用可能不太实际。

第5章

多重插补：基本原理

虽然ML代表了处理缺失数据传统方法的一个重要的发展，但它也有其局限性。如我们已经看到的，关于线性模型和对数线性模型的ML理论和软件容易获取，但超出这两者之外的理论抑或软件通常比较缺乏。例如，如果你想要估计一个Cox比例风险模型或一个有序的logistic回归模型，你将难以对缺失数据执行ML方法。甚至假使你的模型可以用ML估计，但你也可能难以找到你特别需要的、专业的诊断或图形输出软件。

值得庆幸的是，有一个备选方法——即多重插补，它有着与ML相同的最适特性，但却排除了某些局限性。更明确地，当数据为MAR时，正确使用多重插补（MI）会产生一致的、渐近有效且渐近正态的估计值。不同于ML，MI几乎可以被任何一种数据及任何一种模型所使用，且分析可以利用未修改的、传统的软件执行。当然，MI也有它自身的缺点。它的执行可能很麻烦，且也容易出错。这两个问题可以通过使用好的软件做插补来解决。但它最致命的缺点是，每次你使用MI时，它都会产生不同的估计值（但愿差异很小）。这可能导致奇怪的情况，即不同研究者使用相同的方法、相同的数据却得到不同的数字。

第 1 节 | 单一随机插补

MI 不产生一个单一组数目的原因是，故意在插补过程中引入了随机变异。若没有一个随机成分，决定性的插补方法通常对有缺失数据的变量产生低估的方差估值，而且有时候，协方差也一样会被低估。如我们在第 4 章所讲到的，对于多元变量正态模型的 EM 运算法，可以使用残差方差和协方差估计值来矫正传统公式，来解决这个低估方差/协方差的问题。然而，一个好的备选方法是随机抽取自每一个插补变量的残差分布，并把这些随机数字加到插补值。另外，传统公式也可以用来计算方差和协方差。

这里举一个简单的例子。假设我们想要估计 X 和 Y 之间的相关性，但 50% 的个案有 X 缺失数据。我们可以通过对有完整数据的个案回归 X 于 Y，然后再用得到的回归方程产生有缺失 X 的个案的预测值，进而插补缺失的 X 值。作者对一个具 1 万个个案的模拟样本做了这种处理，其中 X 和 Y 抽取自一个标准双变量正态分布中，两者的相关性为 0.30。一半 X 的值被指定为缺失的（完全随机）。利用回归于 Y 的方法来替代缺失值后，X 和 Y 的相关性估计为 0.42。

样本相关性是样本 X 和 Y 的协方差除以它们样本标准差的乘积。为什么利用了回归于 Y 的方法后会高估相关性

呢？首先，回归插补方法产生无偏误的协方差估计值。此外，（没有缺失数据的）Y 的标准差被正确估计为1.0，但（包含插补值的）X 的标准差只有 0.74，而其实际的标准差应为1.0。所以造成了相关度的高估。换一个方法来考虑这个问题，5000 个有缺失数据的个案的 X 插补值为 Y 的一个完全线性函数，因此增大两个变量间的相关性。

我们可以通过从 X 的残差分布随机抽取残差并将这些随机数字加到 X 的预测值上，以矫正这个偏误。在这个例子中，（对 Y 回归的）X 的残差分布为一平均数为 0 且标准差为 0.9525 的正态分布（从成列删除的最小二乘回归估计而得）。对于个案 i，令 μ_i 为一来自标准正态分布的随机抽取且令 $\hat{x}_i$ 为从回归 X 于 Y 而得到的预测值。我们修正的插补值则为 $\tilde{x}_i = \hat{x}_i + 0.9525u_i$。对所有有缺失 X 的观察值而言，我们用 $\tilde{x}_i$ 替代，然后再计算相关性。当作者对这个有 1 万个个案的模拟样本进行这种处理时，（有修正插补值的）X 和 Y 之间的相关性为 0.316，只比真实值 0.300 高一些而已。

第2节｜多元随机插补

随机插补可以消除决定性的插补所特有的偏误，但仍存在一个严重的问题。如果我们把插补的数据(不管是随机或决定性的)当做真实的数据来使用，导致的标准误估计值通常会比较低，而检验统计量会比较高。标准误估计的传统方法不足以说明数据为插补的事实。

对于使用随机插补而言，其解决方案在于不止一次地重复插补过程，从而产生多个完整数据组。因为随机成分，关注的参数估计值在每个插补的数据组中将会只有微小差异。横跨插补的变异可以被用来向上调整标准误。

对于一个有着1万个个案的模拟样本，作者重复随机插补过程8次，产生的估计值见表5.1。虽然这些估计值近似反映无偏误的，但因为没有考虑插补[9]，所以标准误是向下偏误的。我们通过对这8个相关性估计值取平均值，得到一个单一的估计值0.3125。用以下三个步骤来产生改良的标准误估计值：

(1) 估计标准误平方(以得到方差)，并对8个结果取平均值；

(2) 对这8个复制的结果计算相关性估计值的方差；

(3) 加总步骤1和步骤2的结果(运用步骤2中的一个

对方差的小矫正)，并取平方根。

将此置入一个方程中，令 M 为复制的次数，r_k 为复制 k 中的相关性，s_k 为复制 k 中估计到的标准误。则 $\bar{r}$ 标准误的估计值(相关性估计值的平均数)为：

$$S.E.(\bar{r}) = \sqrt{\frac{1}{M}\sum_k s_k^2 + \left(1+\frac{1}{M}\right)\left(\frac{1}{M-1}\right)\sum_k (r_k - \bar{r})^2} \quad [5.1]$$

这个方程可以用于由多重插补所估计的任何参数，r_k 表示所关注参数的第 k 个估计值(Rubin, 1987)。把这个方程运用于我们关注的例子中，得到标准误为 0.01123，大约较由 8 个样本的标准误的平均数高出 24%。

表 5.1 随机插补数据之相关性与标准误

相关性	S.E.	相关性	S.E.
0.3159	0.00900	0.3118	0.00903
0.3108	0.00903	0.3022	0.00909
0.3135	0.00902	0.3189	0.00898
0.3210	0.00897	0.3059	0.00906

第 3 节 | 在参数估计值中考虑随机变异

虽然作者之前描述过的插补缺失数据的方法已经相当完善了，但它并不是最理想的。为了产生 X 的插补值，作者对有完整数据的个案回归 X 于 Y 上，以产生回归方程：

$$\hat{x}_i = a + by_i$$

对于 X 有缺失数据的个案，插补值计算为：

$$\tilde{x}_i = a + by_i + s_{x \cdot y}u_i$$

其中 u_i 是一个来自标准正态分布的随机抽取，而 $s_{x \cdot y}$ 为误差项的估计标准均方根差。对于模拟的数据组，我们得到 $s_{x \cdot y} = 0.9525$。这些值用来对 8 个完整数据组中的每一个产生插补值。

这个方法的问题在于，它视 a、b 和 $s_{x \cdot y}$ 如同真实参数，而非样本估计值。很明显，我们不能得知真实值为多少，但对于“适当”的多重插补(Rubin, 1987)，每一个插补的数据组应该基于不同组的 a、b 和 $s_{x \cdot y}$ 值。这些应该从参数的贝叶斯后验分布中随机抽取。只有这样，多重插补才能完全体现我们关于未知参数的不确定性。

这种观点自然会产生许多疑问。什么是参数的贝叶斯后验分布？我们如何从后验分布中随机抽取用于插补的值？

我们真的需要这额外的复杂程序吗？第一个问题需要另一本书来回答，而且幸好在社会科学的量化运用系列丛书里(Iversen, 1985)有一本相关的书。对于第二个问题，有数个不同的方法可以从后验分布中进行随机抽取，它们有些被包含在容易使用的软件里。在本章稍后部分，当我们考虑在多变量正态模型下的 MI 时，作者将解释一种叫做数据扩增法的方法(Schafer, 1997)。

如果不使用从后验分布中随机抽取的方法可以吗？这个问题的答案很重要，因为有一些随机插补软件，如 SPSS 中的缺失数据模块，可以随机抽取参数值。在许多例子中，作者认为答案是可以的。如果样本够大且有缺失数据的个案比例很小，那么缺少这个额外步骤的 MI 就更容易产生非常接近那些包含这个额外步骤的结果。另一方面，如果样本小或有缺失数据的个案比例很大，则额外的变异可以产生明显的差异。

继续我们所关注的例子。作者用数据扩增法插补 8 个新数据组，以从参数的后验分布中产生随机抽取。表 5.2 提供 X 和 Y 间的相关性及每一个数据组的标准误。相关性估计值的平均数为 0.31288。使用方程 5.1，估计标准误为 0.01329，比由较粗糙的插补方法所获得的 0.01123 稍大。一般而言，当用于插补中的参数为随机抽取时，标准误会略显较大。

表 5.2 使用数据扩增法的随机插补数据之相关性与标准误

相关性	S. E.	相关性	S. E.
0.30636	0.0090614	0.32086	0.0089705
0.31316	0.0090193	0.29760	0.0091143
0.31837	0.0089864	0.32701	0.0089306
0.31142	0.0090302	0.30826	0.0090498

第4节 在多变量正态模型下的多重插补

为了做多重插补,你需要一个模型以产生插补值。对于刚才考虑过的两个变量的例子,作者利用一个具正态分布误差的简单回归模型。很明显,更复杂的情况需要更复杂的模型。然而,对于模型的选择来说,MI 可能没有 ML 那么敏感,因为模型只用来插补缺失数据,而不用来估计其他参数。

理想的情况是,我们会特别建构插补模型以代表每一个数据组的独特的特征。在实际应用中,利用便于使用的、且适用于广泛数据组的、合理良好近似反映的、现成的模型会更加方便。

MI 最受欢迎的模型为多变量正态模型,也就是先前在第4章使用过的、具缺失数据的线性模型的 ML 估计的基础。多变量正态模型意味着:

(1) 所有变量都有着正态分布;

(2) 每一个变量可以用所有其他变量的一个线性函数以及一个正态、同方差的误差项所表示。

虽然这些是很严格的条件,但实际上,即使当有些变量已经明确不是正态分布时,多变量正态模型仍然可以实现好的插补(Schafer, 1997)。对那些没有缺失数据的变量而言,

这是一个完全无害的假设。而对那些有缺失数据的变量而言，正态化转换则可以大大地改善插补的质量。

本质上讲，多变量正态模型下的 MI 是前述两变量例子中使用过的方法的概括。对于每一个有缺失数据的变量，我们估计该变量对于所有其他受关注变量的线性回归。最理想的结果是，回归参数是从贝叶斯后验分布中随机抽取而得。估计到的回归方程接着用来产生有缺失数据的个案的预测值。最后，对每一个预测值，我们加上该变量残差正态分布的一随机抽取值。

插补过程中最复杂的部分是从贝叶斯后验分布中得到随机抽取值。在作者写作本书之时，已有两种用以实现随机抽取的算法可用现有的软件执行：数据扩增法（Schafer，1997）及抽样重要性/重抽样（SIR；Rubin，1987）。这里给出一些执行这些方法的计算机程序。

数据扩增法

NORM　由 Schafer 开发的免费软件，并在其 1997 年出版的书中叙述过。作为一个独立的 Windows 版本或 S-PLUS 的一个程序库，相关资料参见 http://www.stat.psu.edu/～jls/。

SOLAS　一个独立的商业套装，包含数据扩增法（版本 2 及更高版本）和一个倾向评分方法。后者在许多应用上是无效的（Allison，2000）。相关资料参见 http://www.statsolusa.com。

PROC MI　一个 SAS 程序，可于 8.1 或更高的版本中使

用,相关资料参见 http://www.sas.com。

抽样重要性/重抽样

AMELIA 一个由 King、Honaker、Joseph、Scheve 及 Singh(1999)开发的免费套装。作为一个独立的 Windows 程序或 Gauss 的一个模块使用。相关资料参见 http://gKing.harvard.edu/stat.shtml。

SIRNORM 一个由 C. H. Brown 和 X. Ling 写的 SAS 宏命令。相关资料参见 http://yates.coph.usf.edu/research/psmg/web.html。

两种算法理论上都有一定的合理性。SIR 的提倡者(King, Honaker, Joseph & Scheve, 2001)宣称它需要极少的计算机运算时间。然而,这两种方法的相对优越性尚无定论。因为作者有较多关于数据扩增法的经验,因此将会在本章接下来的部分着重介绍这个方法。

第 5 节 | 多变量正态模型的数据扩增法

数据扩增法是马尔可夫链蒙特卡尔运算法的一种形式，一个寻找后验分布的普遍方法，在贝叶斯统计学中越来越受欢迎。在这个部分，作者将叙述它如何处理多变量正态模型。虽然现成可用的软件可自动执行大部分的运算，但对于实际进行过程而言，有一个普遍的理解是有帮助的，特别是当事情出错时。

迭代算法的普遍结构与在前一章叙述过的多变量正态模型的 EM 算法很像，但它还需在两个时点做随机抽取，接下来作者会叙述这点差异。在开始 DA 前，必须选择用于插补过程的一组变量。这组变量很明显应该包含所有有缺失值的变量，以及其他在模型中需要进行估计的变量。同时也应该把不在目标模型中但与具缺失数据的变量高度相关的、或与那些具缺失数据的变量的概率有关的、额外的变量包含进来。

一旦选定变量后，DA 包含下列步骤：

(1) 选择参数的起始值。对于多变样正态模型，参数为平均数和协方差矩阵。起始值可以利用成列删除或成对删除的标准公式得到。有在前一章叙述过的用 EM 算法而获得的估计值则更好。

(2) 用平均数和协方差的当前值来获得方程回归系数的估计值,方程中每一个有缺失数据的变量回归于所有被观察到的其他变量上。对每一种缺失数据的形态都这样处理。

(3) 用回归估计值产生关于所有缺失值的预测值。对每一个预测值,加上一个从该变量残差正态分布中得到的随机抽取值。

(4) 用具观察值和插补值的“完整”数据,利用标准公式重新计算平均数和协方差矩阵。

(5) 从新得到的平均数和协方差的后验分布中随机抽取平均数和协方差。

(6) 用随机抽取的平均数和协方差,回到步骤 2 且继续循环其后的步骤,直到达成收敛。用在最后一次迭代中所产生的插补值构成一个完整数据组。

步骤 5 需要进一步的解释。为了得到参数的后验分布,我们首先需要一个先验分布。虽然可以根据关于这些参数先前的看法,但是通常的做法是使用一个“不提供信息”的先验分布,即包含很少或根本不包含与参数相关的信息的先验分布。我们看看它如何在一个简单的状态下起作用。假设我们有测量某个单一正态分布的变量 Y 的 n 个样本数。样本平均数为 $\bar{y}$,样本方差为 s^2。我们想要从 μ 和 σ^2 的后验分布中随机抽取平均数和方差。利用一个不提供信息的先验分布,[10] 我们可以从自由度为 $n-1$ 的卡方分布中进行抽样、对抽取值取导数、并将结果乘以 ns^2,以得到方差的随机抽取值 $\tilde{\sigma}^2$。接着我们从一平均数为 $\bar{y}$ 且方差为 $\tilde{\sigma}^2/n$ 的正态分布中抽样,以得到平均数的一个随机抽取值。

如果没有缺失数据,这些会是来自真实参数的后验分布

的随机抽取值，但如果我们有插补缺失数据，那么实际上有的值是当插补数据为真实数据时所得到的后验分布的随机抽取值。同样，在给定当前参数值的条件下，当我们随机在步骤 3 中插补缺失数据时，有的值为缺失数据的后验分布的随机抽取值。然而，因为当前值可能不是真实值，插补数据也可能不是真实后验分布的随机抽取值，这也就是为什么程序必须是迭代的。通过持续地在参数的随机抽取（取决于观察到的及插补数据）和缺失数据的随机抽取（取决于当前参数）之间不断来回，我们最终得到从数据和参数两者的联合后验分布中得到的随机抽取值，而这个仅取决于被观察到的数据。

第6节 | 在数据扩增法中收敛

当你运用数据扩增法时，你必须指明反复的次数。然而，这产生了一个难题：需要多少次的迭代才能得到缺失数据和参数的联合后验分布的收敛值？如使用在EM运算法中的最大似然迭代估计，估计值会收敛到单一组的值。接着收敛可以很容易地通过检查从一个迭代到另一个迭代参数估计值改变的大小来进行评估。另一方面，对于数据扩增法，运算会收敛到一个概率分布，而非单一组的值。很难确定收敛事实上是否达成。虽然有些诊断统计量可以用来评估收敛(Schafer, 1997)，但它们根本不可靠。

在大多数运用中，迭代次数的选择都是胡乱瞎猜。关于其可能的范围，为了给出一些提示，Schafer(1997)在他书里的例子中，使用介于50到100之间的迭代次数。次数越多越好，但每一次迭代其运算强度可能很大，特别是对有许多变量的大样本来说。指明一个较大的迭代次数可能会让你长时间痛苦地盯着你的屏幕。

有许多原则要谨记在心。第一，缺失数据(事实上是缺失信息，与缺失数据不大相同)的比例越高，就需要越多次的迭代以达成收敛。如果只有5%的个案有缺失数据，你用小数目的迭代次数就可能过得去。第二，EM运算法的收敛速

度是数据扩增法收敛速度的有效暗示和征兆。一个好的经验法则是，DA 的迭代次数至少要与 EM 所需的迭代次数相同。这也是为什么总在数据扩增法之前进行 EM 的另一个原因(第一个原因是 EM 为数据扩增法提供了良好的起始值)。

作者对于迭代法这个议题的感触是，在大多数的运用中它不是那么重要。从决定性的插补到随机的插补是一个巨大的改进，甚至就算这个随机插补其参数并不是随机抽取的，也已经是巨大的改进。而从没有随机抽取参数的随机插补到有随机抽取参数的随机插补是另一个大的改进，但这个改进毫不起眼。从迭代次数少的数据扩增法到迭代次数多的是更进一步的改进，但在大多数运用中，其边际回报可能相当小。

另一个复杂性来源于多重插补产生多元数据组的事实。至少需要两个数据组，越多越好。给定固定的运算时间，我们可以产生更多数据组，或对每一个数据组产生更多次数据扩增法的迭代。可惜的是，有着很多缺失信息的数据组同时需要更多的迭代和更多数据组。虽然有关这个议题写得很少，但作者更倾向于优先考虑额外的数据组。

第7节｜连续的数据扩增法相对平行的数据扩增法

我们刚才已经看到如何使用数据扩增法来产生单一完整的数据组。对于多重插补,我们需要数个数据组。有两个方法被提出用来执行此任务:

(1) 平行的。对每一个想要的数据组进行一个别系列的迭代。这可以从同一组起始值开始(如 EM 估计值),也可以从不同的起始值开始。

(2) 连续的。进行一长系列的数据扩增法循环。取每第 k 次迭代产生的插补,k 为原来给定数据组的想要的数目。例如,如果我们有五个数据组,我们可以先进行 500 次迭代,再接着使用产生自每第 100 次迭代的插补。在第一次插补前,500 次大数目的迭代形成了一个允许该过程收敛到正确分布的测试时期。

两种方法都可以接受。连续方法的一个优点是,较容易收敛到真实后验分布,特别是对于那些在序列中位置较后的数据组。然而,每当你从相同的一系列迭代中取出多个数据组时,却不能确定那些数据组是否在统计上相互独立,而这种独立性是有效推论的必需条件。在相同系列中两个数据组越接近,它们就越有可能存在某些依赖性/相关性。这也

就是为什么你不能只运行 200 次迭代以得到收敛，并且接着用 5 次迭代来产生五个数据组。

平行方法避免了依赖性/相关性的问题，但更不能确定是否达到收敛。此外，Rubin(1987)和 Schafer(1997)都建议与其对每一个序列都使用相同组的起始值，不如从一个以 EM 估计值为中心的“过度分散的”先验分布中抽取起始值，但这并不总是容易执行。[11]

对于大量广泛的运用来说，作者认为选择连续或平行方法不会有重大差异。若有着同样次数的迭代，两种方法应该会给出几乎相同的结果。作者相信在大多数的例子中，当使用平行方法时，以 EM 估计值为每一个迭代系列的起始值，都能获得可接受的结果。

第8节 | 对非正态或类别数据使用正态模型

可以以多变量正态分布紧密近似反映的数据组事实上很稀有,有时会出现具高度偏态分布的变量以及其他完全是类别型的变量。在诸如此类的例子中,我们刚才考虑过的根据正态的方法有任何的价值吗? 如稍早前叙述过的,对于没有缺失数据的变量都没问题,因为它们没有缺失数据的变量而且不需要被插补。

对于有缺失数据的变量,大量的证据显示这些插补方法可以处理得相当好,甚至当分布明显不是正态的时候(Schafer,1997)。然而,有一些技巧可以改善插补非正态变量的正态模型的绩效。

对于高度偏态量化变量,在执行插补前先转换这些变量以降低偏度,通常是有帮助的。任何可以胜任这项任务的转换应该都是可行的。在数据已经被插补后,相反的转换可以运用来将该变量变回其原始的度量标准状态。例如,对数转换能大量降低大多数收入数据的偏度;在插补后,只要取收入的反对数即可。这对于有限制范围的变量帮助特别大。如果你插补收入的对数而非收入本身,就不可能产生小于0的收入插补值。同样地,如果要插补的变量是比例,logit 转

换就可以防止大于 1 或小于 0 的插补。

有些软件可以用另一种方法处理有限制范围的问题。如果你对一个特别变量指明一个最大或最小值，它将会拒绝所有在这个范围外的随机抽取值，并简单地做额外的抽取直到它处于指定范围中。虽然这是一个很有用的可选方法，但使用转换以降低变量中的偏度仍是令人满意的。

对于离散的量化变量，通常需要把插补值做合适的四舍五入以变成一个离散的量尺。例如，假设成年人被问及他们有多少个小孩。对这个问题的回答，其分布会是典型的偏态，所以可以通过运用对数或从平方根转换开始。在插补后，反向转换将会产生非整数的值。这些可以被四舍五入取整数以符合原始量尺。有些软件可以自动执行诸如此类的四舍五入。

若完全是类别的变量又怎样呢？虽然有方法和计算机程序设计仅用于只有类别变量的数据组，以及有类别的和正态分布变量的混合的数据组，但这些方法更加难以使用且通常会彻底地失效。许多使用者也会运用有较少修改的正态模型来执行。二分类的变量，如性别，通常使用有着 0 或 1 的虚拟（指标）变量来代表。任何对一个二分类变量的转换将仍产生一个二分法，因此没有值可以用来试着降低偏度。相反，我们可以如同其任何其他变量一样简单地插补该 0—1 变量。接着根据该插补值是否高于或低于 0.5，再四舍五入插补值至 0 或 1。大部分的插补会落于（0，1）这个区间内，有时候仍会落在该区间外。在这个例子中没有问题，因为我们根据插补值较接近 0 或 1 者，来指派 0 或 1 的值。

如果有超过两个类别的变量，通常由虚拟变量组来表

示。在数据扩增法阶段不需要做任何特别的事,但当指派最终值时则需要小心。问题是我们需要指派个体至一个类别且只有这么一个类别,并适当编码所有的虚拟变量。假设被插补变量为婚姻状态,有三个类别:从未结婚、目前已婚、曾结过婚。令 N 为从未结婚的虚拟变量,令 M 为目前已婚的虚拟变量。用这两个变量做插补且使用插补值产生最终的编码。这里有一些可能的插补和得到的编码:

插补值			最终值	
N	M	$1-N-M$	N	M
0.7	0.2	0.1	1	0
0.3	0.5	0.2	0	1
0.2	0.2	0.6	0	0
0.6	0.8	−0.4	0	1
−0.2	0.2	1	0	0

基本原则是这样的。除了两个插补值外,也要计算1减去这两个插补值的总和,这可被视为参照组的插补值。接着确定哪一个类别有最高的插补值。如果该值对应着一个明确为虚拟变量的类别,则指派1给该变量。如果最高的值对应参照组,则指派0给另外两个虚拟变量。此外在这个背景下负值可能显得比较奇怪,但此法仍然可以运用。延伸至四个或更多个类别应该就很容易理解且简单易做了。

第 9 节 | 探索分析

很多数据分析包含探索工作，在其中分析实验的各种方法和模型。对于任何已经做过这种工作的人而言，多重插补的过程似乎问题比较大。对数个数据组同时执行探索分析必然是一个繁琐的过程。此外，对每个数据组的分析也可能建议使用有细微差异的模型，但多重插补对所有数据组要求一个相同的模型。

解决方法很简单，但却显得特别随意。当产生多元数据组时，只要产生比你做多重插补分析时所需要的多一个的数据组即可。因此，如果你想要做三个数据组的多重插补分析，就会产生四个数据组。之后再用这个额外的数据组做探索分析。一旦你决定某单一模型或小组模型，就要对剩余的数据组重新估计这些模型，并运用我们已经讨论过的结合结果的方法。要谨记，虽然从探索分析获得的参数估计值将近似反映无偏误的，但是所有的标准误会向下偏误且统计检定量会向上偏误。因此，使用更加保守的准则而非通常（有着完整数据）的准则来衡量一个给定模型的适当性可能结果会更令人满意。

第10节 | MI实例1

我们现在有足够的背景知识可以考虑一个现实中的多重插补的例子。让我们重新回顾一下第4章中使用过的例子,一个包含7个变量测量的1302所美国大专院校数据组,其中除了一个变量外,所有的变量都有缺失数据。和之前一样,我们的目标是估计一个预测GRADRAT的线性模型,GRADRAT为高年级毕业生对于四年前以新生身份就读的数目之比率。自变量包含除了ACT(即ACT分数之平均数)以外的所有变量。这个变量(ACT)被包含于插补过程中以得到较佳的CSAT(即结合SAT分数的平均数)预测。后面的变量(CSAT)40%的个案有缺失数据,但对于同时有两个变量(CSAT和ACT)数据呈现的488个个案,CSAT与ACT高度相关($r = 0.91$)。

第一个步骤要检查变量的分布以检验正态性。直方图和正态概率图显示除了一个变量外,所有变量都合理接近正态分布。这个例外是高度左偏的就学这个变量。如同在ML实例中,作者使用就学的自然对数,其分布有着很小的偏度。

为了执行数据扩增法,作者使用SAS的PROC MI。第一步是要用EM运算法来估计平均数、标准差和相关性,结果已在表4.4中呈现。EM运算法用了32次迭代达到收敛。

这是一个不算太大的数目，其或许反映了某些变量有大比例的缺失数据。然而，它还没有大到意味着运用 EM 运算法或数据扩增法会有严重问题。

这里有一个最小组的 SAS 命令语句以产生多重插补：

```
proc mi data = college out = collimp;
    var gradrat csat lenroll private stufac rmbrd act;
run;
```

college 为输入数据组（对于缺失值用点表示）的名称，而 collimp 为输出数据组（包含被观察到的值与插补值）的名称。var 述句给予用于插补过程中的变量的名称。PROC MI 的默认值为根据一系列连续性的迭代产生的五个完整数据组。以 EM 估计值作为起始值，在第一次插补前有 200 次“测试”迭代。接着有 100 次在连续的插补之间的迭代。五个数据组被写入一个大的 SAS 数据组以利于后面的分析。输出数据组包含一个新的变量_imputation_，其值由 1 到 5 表示不同的数据组。因此，原始数据组有 1302 个观察值，新数据组有 6510 个观察值。

不采用默认值，作者实际上采用了一个稍微复杂的程序：

```
proc mi data = my.college out = collimp seed = 1401;
    minimum = 0600..0126011
    maximum = 1001410..100870031
    round = 1111.111;
    var gradrat csat lenroll private stufac rmbrd act;
    MCMC nbiter = 500 niter = 200;
run;
```

seed = 1401 对于随机数字产生器设定一个种子值，以至于结果可以在后面一次执行中确实能被重新产生。maximum

和 minimum 选择设定对每一个变量的最大和最小值。如果一个随机插补值恰巧落在这些边界外,则该值会被拒绝,而一个新的值会被抽取。对 gradrat 和 stufac,最大值和最小值的理论边界为 0 和 100。对于 lenroll 和 private,没有指明边界。对于 csat、rmbrd 和 act,作者使用每一个变量观察到的最大值和最小值。round 选择四舍五入所有变量的插补值至整数,除了 STUTAC 以外,它被四舍五入到小数点后第一位。

MCMC 述句允许对数据扩增法过程采取更多控制。作者在此已指明在第一次插补前要有 500 次测试的迭代,接着在连续的插补之间有 200 次迭代。

这些迭代能够足以达到收敛吗?答案很难确定,但我们可以试验由 Schafer(1997)建议的一些收敛诊断。一个简单的做法是检查产生自每一个迭代的一些参数值,并看所有迭代得到的结果之间是否有任何趋势。对于有七个变量的多元正态模型,参数为 7 个平均数、7 个方差和 21 个协方差。不检查所有参数,只专注于那些牵涉有最多缺失数据的变量的参数,因为这些牵涉有最多缺失数据的变量的参数最有可能出现问题。对于这些数据,变量 CSAT 有 40%的缺失数据,因为终极目标是要估计预测 GRADRAT 的回归,因此我们看一下 CSAT 的二变量回归斜率,也就是 CSAT 和 GRADRAT 的协方差除以 CSAT 的方差。图 5.1 画出了数据扩增法前 100 次迭代的回归斜率值。在第一次迭代后,在斜率系数估计值中似乎没有特别的趋势,这也让人比较放心。

另一个被提议使用的诊断方法,是对所关注的参数于连续迭代中的许多滞后值做一组自相关。目标是要在插补间有足够的迭代,从而使自相关为 0。使用全部系列 1300 个迭

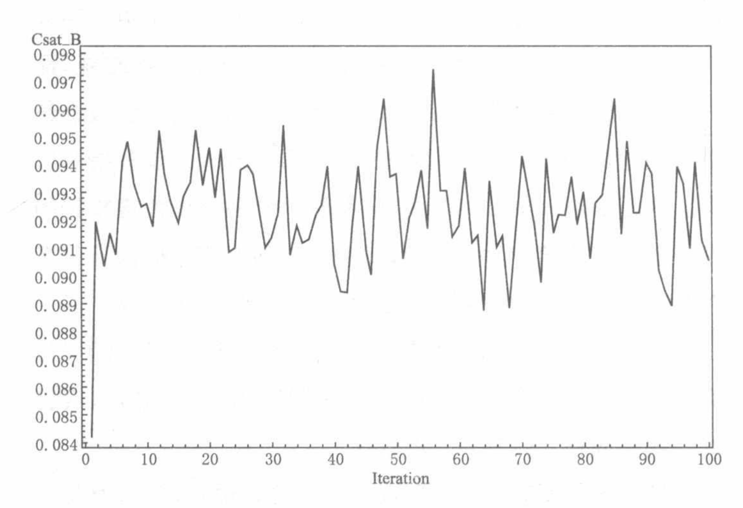


图 5.1 前 100 数据扩增迭代的回归中变量 CSAT(对 GRADRAT)的斜率估计值分布情况

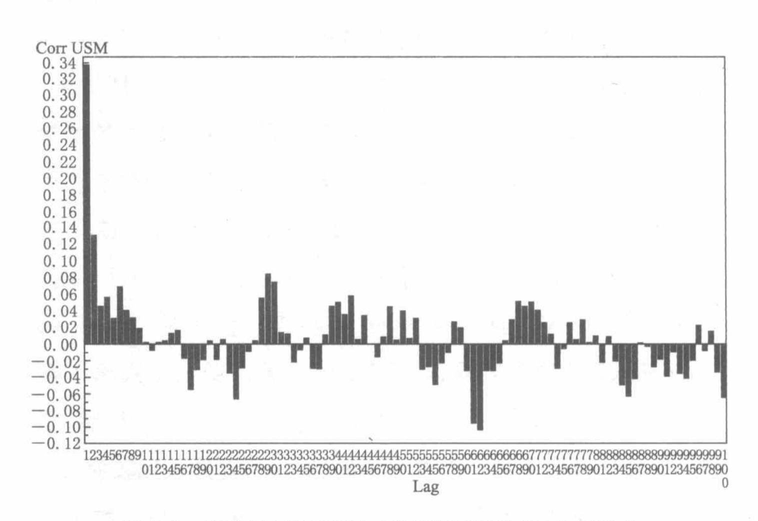


图 5.2 CSAT 对 GRADRAT 的回归斜率的第 1 至 100 个滞后值间变化之自相关

代，图5.2画出了具不同滞后值的二变量回归斜率值间的自相关，因此最左边的值0.34代表距离为一个迭代远的参数值间的相关性。第二个值为两个迭代分开的参数值间的相关性。虽然这两个初始值比较高，但自相关迅速降低至相当低的值(落在0的0.10间)，明显有一个随机的形态。这两个诊断结合一起，建议我们在插补间可以使用远少于200次的迭代。多做一些迭代没有什么不利，而且这里使用的诊断不保证会达到收敛。

在产生完整数据组后，作者可以将转换过的变量就读(ENROLLMENT)的对数转换回其原始形式，但因为作者期望就读(ENROLLMENT)对于毕业率的效应有逐渐降低的回报(RETRUNS)，因此作者决定让这个变量保留对数形式，就如同作者在第4章对以ML估计的回归模型一样。

表5.3 五个完整数据组的回归系数(及标准误)

截 距	CSAT	LENROLL	PRIVATE	STUFAC	RMBRD
−33.219	0.069	1.550	11.632	−0.145	2.951
(4.272)	(0.004)	(0.534)	(1.124)	(0.083)	(0.390)
−33.230	0.067	2.023	12.840	−0.116	2.417
4.250	(0.004)	(0.526)	(1.126)	(0.082)	(0.392)
−31.256	0.071	1.852	12.274	−0.213	1.657
(4.306)	(0.004)	(0.546)	(1.157)	(0.084)	(0.408)
−34.727	0.068	2.187	13.468	−0.142	2.103
(4.869)	(0.004)	(0.532)	(1.121)	(0.083)	(0.391)
−29.117	0.065	1.971	12.191	−0.231	2.612
(4.924)	(0.004)	(0.538)	(1.141)	(0.084)	(0.393)

所以下一个步骤就是简单地对五个完整数据组中的每一个估计回归模型。通过使用SAS中的BY数据可以使执行变得更容易，从而避免指明五个不同的回归模型：

```
proc reg data = college outset = estimate covout;
    model gradrat = csat lenroll private stufac rmbrd;
    by_imputation_;
run;
```

这组数据告诉 SAS,要对由五个_imputation_值所定义的每一个次群体分别估计一个回归模型。outset = estimate 要求回归估计值被写入一个叫做 estimate 的新数据组,而 covout 要求回归参数的协方差矩阵被包含进该数据组中。这使得在下一步骤中结合估计值显得容易。五个回归结果显示于表 5.3 中。很显然,每一个回归到另一个回归具有很大的稳定性,但也有值得注意的变异,归因于插补的随机成分。这些回归的结果用另一个叫做 MIANALYZE 的 SAS 程序,整合成一个单一估计值组。由下列数据引起:

```
proc mianalyze data = estimate;
    var intercep csat lenroll private stufac rmbrd;
run;
```

这个程序直接对包含有回归运作所产生的系数和相关统计量的数据组estimate来起作用。结果呈现于图 5.3。

Multiple-Imputation Parameter Estimates

Variable	Mean	Std Error Mean	DF	t for H0: Mean=0	Pr>|t|	Fraction Missing Information
intercept	−32.309795	5.639411	72	−6.596995	<0.0001	0.255724
csat	0.068255	0.004692	39	14.547388	<0.0001	0.356451
lenroll	1.916654	0.595229	110	3.220027	0.0017	0.206210
private	12.481050	1.367858	40	9.124524	<0.0001	0.344151
stufac	−0.169484	0.099331	42	−1.706258	0.0953	0.329284
rmbrd	2.348136	0.670105	10	3.504132	0.0067	0.708476

图 5.3 从 PROC MIANALYZE 得出的部分结果

图 5.3 中标示为“平均数”的字段包含表 5.3 中系数的平均数。使用方程 5.1 计算的标准误明显比表 5.3 中的标准误大，因为不同回归间的变化被加到回归内的变化了。然而，对于一些系数而言，有比其他系数更多的回归间的变化。在低端，lenroll 系数的标准误在图 5.3 中只比表 5.3 的标准误平均值大了约 10%。在高端，rmbrd 的结合标准误较个体标准误平均值大了约 70%。在表 5.3 中明显可以看出 rmbrd 系数具较大的变异，其估计值为 1.66 到 2.95。

图 5.3 中标示为“t for H0：Mean = 0”的字段就只是每一个系数对于其标准误的比。紧接着的字段给出了用于从 t 表中计算 p 值的自由度。这个数目与观察值的数目或变量的数目无关。虽然没必要知道自由度是如何计算的，但作者认为需要对该值做简短的解释。对于一个给定的系数，令 U 为回归模型中标准误平方的平均值。令 B 为回归间系数的方差。因缺失数据导致方差中相对的增加则被定义为：

$$r = \frac{(1 + M^{-1})B}{U}$$

其中 M 和之前相同，是用来产生估计值的完整数据组数目。自由度则被计算为：

$$df = (M - 1)(1 + r^{-1})^2$$

因此，相对于回归内变化回归间变化越小的话，自由度越大。有时候，自由度将会远大于观察值的数目。但毋需担心，因为任何大于 150 左右的数目会造成一个实质上如同依标准正态分布的 t 表。然而，有些软件（包含 PROC MIANALYZE）可以产生一个不大于样本数目的、调整过的自由度

(Barnard & Rubin, 1999)。

最后一栏缺失信息比是因缺失数据导致每个系数有多少信息缺失的估计值,其范围比 lenroll 的低 21%,高于 rmbrd71%。对于有着 40%缺失数据的 rmbrd 而言,有高的缺失信息并不令人吃惊,但对于没有缺失数据的 private 及少于 1%缺失数据的 stufac 而言,缺失信息则高得令人吃惊。要理解这点,必须要先知道一些知识。首先,对于每一个给定系数,其缺失信息量不只取决于该特定变量的缺失数据,也取决于与其相关的其他变量之缺失数据百分比。其次,MI-ANALYZW 程序没法知道每一个变量有多少缺失数据。相反,缺失信息估计值全然立足于回归内和回归间的相对变化。如果在回归间有大的变化,则暗示有较多缺失信息。有时候缺失信息比以 γ 表示,用我们刚定义过的两个统计量 r 和 df 计算,即:

$$\hat{r} = \frac{r + 2/(df + 3)}{r + 1}$$

同时我们要明确,在表中报告的缺失信息比可能只是一个受到大量抽样变化所影响的估计值。

如前所述,多重插补的麻烦之一是它不会产生一个确定的结果。每次执行多重插补,都会得到有些微差异的估计值和相关的统计量。为了理解这点,我们看一下图 5.4,其立足于由一个全新的数据扩增法运算所产生的 5 个数据组。虽然 lenroll 和 private 的缺失信息比之前的低很多,但大多数的结果与图 5.3 非常相似。

当缺失信息比较高时,多于平常建议的三到五个完整数据组可能是有必要的,从而得到稳定的估计值。那么到底需

Multiple-Imputation Parameter Estimates

Variable	Mean	Std Error Mean	DF	t for H0: Mean=0	Pr>|t|	Fraction Missing Information
intercept	−32.474158	4.816341	124	−6.742496	<0.0001	0.192429
csat	0.066590	0.005187	20	12.838386	<0.0001	0.489341
lenroll	2.173214	0.546177	2157	3.978955	<0.0001	0.043949
private	13.125024	1.171488	1191	11.203719	<0.0001	0.059531
stufac	−0.190031	0.099027	51	−1.918988	0.0607	0.307569
rmbrd	2.357444	0.599341	12	3.993396	0.0020	0.623224

图 5.4 从 MIANALYZE 多重插补复制而得的输出结果

要多少数据组呢?一个有着无限数目数据组的多重插补为完全有效(如同 ML),但具有限数目数据组的 MI 不能达到全然有效。Rubin(1987)证明了一个根据 M 个数据组的估计值与一个根据无限数目数据组的估计值,两者之相对有效性为 $(1+\gamma/M)^{-1}$,其中 γ 为缺失信息比。这暗示有五个数据组和 50%的缺失信息,其估计程序的效率为 91%。有 10 个数据组的话,效率提升至 95%。同样,只有五个数据组得到的标准误比由无限数目的数据组所提供的标准误大 5%。10 个数据组产生的标准误比由无限数目的数据组所提供的标准误大 2.5%。其结果是,即使有着 50%的缺失信息,但五个数据组已表现得相当好了。将数据组数目扩大至其两倍,可以将过量的标准误减半,但此时的标准误已经很小就不需要再这样做了。

在结束回归实例前,让我们比较图 5.4 的 MI 结果和表 4.6 的 MI 结果。我们发现,系数估计值相当相似,标准误和 t 统计量也一样相当相似。毫无疑义,这两个分析会得出相同的结论。

第6章

多重插补：复杂化

第 1 节 | MI 中的交互作用和非线性

虽然我们之前叙述过的方法对于估计有缺失数据的变量的主要效应有非常好的效果，但它们对于估计交互作用效果可能并不理想。例如，假设我们怀疑公立和私立大专院校，SAT 分数(CSAT)对于毕业率(GRADRAT)的效应有所不同。一个检验这个假设的方法(方法 1)是取之前插补的数据，创造一个等于 CSAT 和 PRIVATE 乘积的新变量，并将这个新变量与已在模型中的其他变量一起纳入回归方程中。

表 6.1(方法 1)最左边的一栏显示了这种做法的结果。变量 PRIVACAST 是 CSAT 和 PRIVATE 的乘积。p 值为 0.39，交互作用在统计上一点儿也不显著，所以我们可以总结为，在公立和私立机构间，CSAT 的效应没有改变。

表 6.1　有交互作用项的回归：三种方法

变　量	方法 1		方法 2		方法 3	
	系数	p 值	系数	p 值	系数	p 值
INTERCEPT	−39.142	0.000	−48.046	0.000	−50.2	0.000
CSAT	0.073	0.000	0.085	0.000	0.085	0.000
LENROLL	2.383	0.000	1.932	0.001	1.950	0.013
STUFAC	−0.175	0.208	−0.204	0.083	−0.152	0.091
PRIVATE	20.870	0.023	35.128	0.001	36.118	0.002
RMBRD	2.134	0.002	2.448	0.000	2.641	0.003
PRIVCSAT	−0.008	0.388	−0.024	0.022	−0.024	0.024

这个方法的问题是，虽然多元正态模型精于插补新产生变量间线性关系的值，但它并不模型化任何其他高阶动差。因此，除非使用特殊技巧，不然插补值不会显示出具有交互作用的证据。在这个例子中，交互作用两个变量中的其中一个(PRIVATE)是二分的，最自然的解法(方法 2)是对私立大专院校和公立大专院校分别做系列的数据扩增法。这允许CSAT 和 GRADRAT 间的关系在这两个群体间有所差异，且允许插补值反映这个事实。一旦完成分别的插补，数据组会重新结合成一个单一数据组，乘积变量被创造，就可用该乘积变量来运行回归。表 6.1 中间部分的结果显示，PRIVATE 和 CSAT 间的交互作用在 0.02 程度上显著。我们可以更明确地发现，CSAT 对于毕业率的正效应，私立大专院校比公立大专院校小。

第三个方法(方法 3)对所有包含 CSAT 和 PRIVATE 的观察值的个案，在插补前创造乘积变量，接着再如同任何其他有缺失数据的变量一样插补该乘积变量，最后使用插补数据估计包含该乘积变量的回归模型。

这个方法不如方法 2 那么吸引人，因为很明显乘积变量会有一个根本不是正态分布的，而在插补过程中却假设了正态性。然而，如同在表 6.1 右边部分所看到的，方法 3 的结果与方法 2 的结果非常接近，而且无疑比方法 1 更接近方法 2 的结果。

方法 3 的结果令人放心，因为当交互作用中的两个变量都被以一量化量尺测量时，方法 2 并不可行。因此，如果我们希望估计一个有 CSAT 和 RMBRD 的模型，我们需要对具这两个变量数据的 476 个个案创造一个乘积变量。对于其

他剩余的 826 个个案，我们必须插补这个乘积变量作为数据扩增法过程的一部分。想估计一个牵涉具缺失数据的变量之非线性关系的模型时，就应该使用这个方法（或方法 2，当其为可行时）。例如，如果我们想要估计一个有 RMBRD 和 RMBRD 平方项两者的模型，则平方项必须被插补作为数据扩增法过程的一部分。这个需求会对在开始插补前就期待想要的函数形式的插补者造成一些负担。这也表示，我们必须从已被其他人使用的完全线性模型进行插补的数据中，谨慎地估计非线性模型。当然，如果一给定变量的缺失数据比例较小，我们选择用一个变量的原始形式插补它，随后再建构一个非线性的转换，可能会有效果。毫无疑义，对于 1302 个个案，只有两个个案有缺失数据的 STUFAC 变量（学生与教职员比例），在将两个插补值平方后，将 STUFAC 平方项放入回归模型是可以接受的。

第2节 插补模型和分析模型之适合性

交互作用这个问题说明了一个在多重插补中更普遍的议题。比较理想的结果是,用来插补的模型与用于分析的模型一致,而且两者都正确地代表该数据。计算标准误基本公式(方程5.1)取决于它的适合性和正确性。

如果插补和分析的模型不同时会怎么样?这取决于差异的本质以及哪一个模型更加正确(Schafer, 1997)。特别要关注的是那些在一个模型中为另一个模型特例的情况。例如,插补模型可以考虑交互作用,但分析模型可能不行,或者分析模型可以允许交互作用,但插补模型可能不行。这两个例子的其中任意一个,如果被较简单模型所施加的额外限制为正确的,那么我们已讨论过的在多重插补下的推论程序就是有效的。然而,如果额外的限制不正确,那么使用标准方法的推论可能就无效。

对于模型选择较不敏感的方法也被提议用来估计多重插补下的标准误(Wang & Robins, 1998; Robins & Wang, 2000)。很明显,当插补和分析的模型不兼容或者当两个模型都不正确时,这些方法给予有效的标准误估计值。然而,在任一阶段的不正确模型都仍有可能产生偏误的参数估计值,而且备选方法需要目前尚不能获取的专业软件。

第3节 | 插补中因变量所扮演的角色

因为GRADRAT是包含在数据扩增法过程中的变量之一，于是因变量暗中被用来插补自变量的缺失值。这样合理吗？它不会导致产生虚假的大的回归系数吗？答案是这不仅可行，而且为了要得到无偏误的回归系数估计值，这甚至是必要的。决定性的插补，使用因变量以插补自变量的缺失值，的确可能产生虚假的大的回归系数，但将一个随机成分引入插补过程中，就可以抵消这个倾向并给予我们近似反映无偏误的估计值。事实上，将因变量排除于插补过程外，对于那些有缺失数据的变量而言，至少会倾向产生虚假的小的回归系数(Landerman, Land & Pieper, 1997)。在大专院校的例子中，如果GRADRAT不被用于插补，就会有大部分缺失数据的CSAT和RMBRD，两者的系数分别会降低约25%和20%。同时，只有5个缺失值的LENROLL的系数会变大65%。

当然，将GRADRAT纳入数据扩增法过程中也意味着GRADRAT的所有缺失值被插补了。有些学者就反对对因变量插补缺失值(Cohen & Cohen, 1985)。鉴于此项建议，我们需要在开始插补前，就去掉任何在因变量上有缺失数据的个案。这个建议有一个正当的基本理由，但它只在一些特

别的例子中适用。如果因变量有缺失数据但所有的自变量没有缺失数据，则（不论线性或非线性的）回归模型之最大似然估计没有使用具缺失数据的个案的任何信息。因为ML是最合适的，在多重插补下插补缺失个案并不会获得任何好处。事实上，虽然诸如此类的插补不会导致任何偏误，但其标准误会比较大。然而，当自变量上也有缺失数据时，情况就不同了。因变量上有缺失数据的个案，确实有一些可以用于回归系数估计的信息，虽然信息量可能不是很大。结果是，在因变量和自变量两者皆有缺失数据的典型例子中，因变量上有缺失数据的个案不应该被删除。

第 4 节 | 在插补过程中使用额外的变量

如之前所提到的，用于数据扩增法的变量组确实应该包含那些将在计划分析中使用到的所有变量。在大专院校的例子中，我们也纳入一个额外变量 ACT(平均 ACT 分数)，因为其与有着大量缺失数据的变量 CSAT 高度相关。目的是要改善 CSAT 的插补以得到它更可靠的回归系数的估计值。如果我们包含其他与 CSAT 相关的变量，我们会做得更好。

一个相对较简单的例子就能阐明额外预测变量的好处。假设我们想要估计 1302 所大专院校的 CSAT 平均分数，我们知道，有 523 个个案有 CSAT 缺失数据。如果我们使用有 CSAT 值的其他 779 个个案计算平均数，我们会得到表 6.2 第一行的结果。第二行则显示用多重插补及变量 ACT 的估计平均数(及标准误)。平均数降低了 9 个点，而标准误降低了 13%。虽然 ACT 和 CSAT 的相关性约为 0.90，但它作为一个预测变量的有用性，却被 523 个缺失 CSAT 的个案中只有 226 个个案有 ACT 的观察值这个事实所影响了。如果我们加入一个额外的变量 PCT25(班上前 25%的学生的比例)，我们的标准误会进一步减少。PCT25 与 CSAT 的相关性约为 0.80，且对于额外的 240 个 CSAT 和 ACT 两者皆缺失数据的个案，PCT25 更容易获取。

表 6.2　不同变量用于插补中所得的 CSAT 之平均数(及标准误)

用于插补中的变量	平均数	标准误	缺失信息%
没有	967.98	4.43	40.1*
ACT	956.87	3.84	26.5
ACT, PCT25	959.48	3.60	13.3
ACT, PCT25, GRADRAT	958.04	3.58	11.3

注：* 缺失数据的实际百分比。

表 6.2 最后一行加入了与 CSAT 相关性约为 0.60 的 GRADRAT,但只能利用 17 个未被 PCT25 或 ACT 所涵盖的个案。丝毫不出人意料,标准误的减少相当小。当作者试着引入图 5.4 中回归模型的所有其他变量时,标准误事实上变得更大。这可能是由于其他变量与 CSAT 的相关性低很多,然而因为需要估计他们的回归系数以预测 CSAT,所以引进额外的变量。如同其他预测问题一样,当粗劣的预测量被加入模型时,插补可能会变得更糟。

第 5 节 | 多重插补的其他参数方法

如我们已经看到的，在广泛的数据类型和缺失数据形态下，多变量正态模型下的多重插补是相当简单的。作为处理缺失数据的一个惯例方法，这可能是当前最好的方法。然而，有数个备选方法在某些情况下可能更为可取。

多元正态模型最明显的局限性之一是，它只是被设计用来插补量化变量的缺失值。如我们已经看到的，类别变量通过使用一些临时的方法来修正。然而，有时候你可能想做得更好。对于在插补过程中所有变量都是类别变量的情况下，有一个更好的模型，即无限制的多项模型(其在列联表中每一个单元格都有一个参数)或者一个允许在多项参数上有所限制的对数线性模型。在第 4 章中，我们讨论过这些模型的 ML 估计。Schafer(1997)也证明了这些模型如何能被使用以作为数据扩增法的基础，从而产生多重插补，而且他也开发了一个叫做 CAT 的免费软件程序以执行这个任务(参见 http://www.stat.psu.edu/~jls/)。

当数据包含类别的和量化变量的组合时，另一个 Schafer 程序(MIX)使用数据扩增法以产生插补。这个方法假设类别变量有一多项的分布，并在变量上可能有对数线性的限制。在由类别变量所创造出的列联表中的每一个单元格里，

量化变量被假设有一个多变量正态分布。这些变量的平均数可以在各单元格变化,但协方差矩阵则被假设为固定的。

在作者写这本书时,CAT 和 MIX 还只能从 S-PLUS 统计软件包的程序库中获取,但它们之后可能会推出独立的版本。在之前的两个例子(CAT 和 MIX)中,基准模型包含的参数远远超过一般多元模型所包含的参数数目。因此,这些方法的有效运用很明显需要更多来自插补执行者的知识及创造,以及更大的样本数以达到稳定的估计值。

如果数据对于一个单一类别变量有缺失,logistic 回归模型下的多重插补就相当容易(Rubin, 1987)。假设在编码为五个类别的婚姻状态数据有缺失,且有数个可能的连续的和类别的预测变量。为了插补,我们使用具完整数据的个案,以一个预测量的函数,对婚姻状态估计一个多项的 logit 模型。这将会产生一组系数估计值 $\hat{\beta}$ 及一个协方差矩阵估计值 $\hat{V}(\hat{\beta})$。为了考虑参数估计值间的变异,我们从一个平均数为 $\hat{\beta}$ 且协方差矩阵为 $\hat{V}(\hat{\beta})$的正态分布中随机抽取参数估计值(Schafer 给出了如何有效实行这种操作的实用建议)。对于有缺失数据的每一个个案,抽取的系数值和被观察到的协变量值被代入多项的 logit 模型以产生落入 5 个婚姻状态类别的预测概率。根据这些预测概率,我们随机抽取 1 个婚姻状态类别,作为最终的插补值。[12]这整个过程被重复多次以产生多个完整数据组。当然,一个二元变量只是这个方法的一个特殊例子。这个方法可被用于许多其他的参数模型,包含泊松回归及参数的失效时间回归。

第6节 | 无参数及部分参数方法

相对于我们刚考虑过的全然参数方法，在相对不严格假设的情况下，有人提议用另外几种方法来做多重插补。在这个部分，作者将考虑一些比较有代表性的方法，但需要注意的是，这些方法中几乎每一个都会存在许多变化情况。虽然当数据缺失为单调形态时，这些方法通常可以毫无困难地被普遍化到具多元变量的情况(在第4章已叙述过了)，但这些方法在只有一个单一变量有缺失数据的情况下被运用最为自然。关于该内容可以参考 Rubin(1987)有关单调广义化的内容。当缺失数据不服从一个单调形态时，这些方法有时可以被使用，但在这样的设定下，它们明显缺少坚实的理论基础。

当在参数和非参数方法间选择时，通常要在偏误与抽样变异间有一个权衡。参数方法倾向于有较少的抽样变异，但如果参数模型对于所关注的现象并不是一个良好的近似反映时，它们可能会导致有偏误的估计值。非参数模型在许多情况下可能相对不容易出现偏误的情况，但估计值通常有更多的抽样变异。

热卡方法

最广为人知的非参数插补方法，是美国人口调查局经常

使用的、同时也供公众使用的数据组插补值的“热卡”方法。

其基本理论是,我们想要插补某一特别变量 Y 的缺失值,该变量有可能是量化的或类别的。我们需要寻找一组与 Y 相关的、(没有缺失值的)类别变量 X。我们根据 X 变量制作一个列联表。如果在该列联表某个特定的单元格中个案有缺失 Y 值,我们取在同一个单元格中的一个或多个无缺失的个案,并用它们的 Y 值来插补这个缺失的 Y 值。

很明显,该方法可能会比较复杂,而且最重要的问题是如何选择“捐赠者”值以指派给有缺失值的个案?此外,捐赠者个案的选择应稍微被随机化从而避免偏误。这自然地导致多元回归,因为任何随机化方法可能不止一次地被运用以产生这个缺失值的不同的插补值。诀窍在于要做到随机化,使所有自然的变异都能被保留。为了达到这个目的,Rubin 提出了一个他创造的近似反映贝叶斯自举法(Rubin, 1987; Rubin & Schenker, 1991)。下面具体介绍如何使用这种方法。假设在列联表一特定的单元格中有 n_1 个个案于 Y 有完整数据且 n_0 个个案在 Y 有缺失数据。根据这些步骤:

(1) 从有完整数据的 n_1 个案组中,取一个 n_1 个案的随机样本(有取代);

(2) 从这个样本中,取一个 n_0 个案的随机样本(有取代);

(3) 指派 n_0 个的观察值给有缺失 Y 数据的 n_0 个个案;

(4) 对列联表中每一个单元格重复步骤 1 到 3。

当运用到列联表中所有的单元格时,这四个步骤会产生一个完整的数据组。对于多重插补,整个过程会被重复多次。在每一个数据组中都执行了所要的分析,用的同样是多变量正态插补的公式,同时所得结果加以汇总。

虽然我们似乎可以忽略步骤 1 而直接从有完整数据的 n_1 个案中选取 n_0 个捐赠者个案，但这无法估计标准误产生足够的变异。其他的变异是因为步骤 2 抽样用的是替代的方法。

预测均数匹配

热卡插补一个主要的吸引力在于，所有插补值都是实际被观察到的值。因此，没有“不可能出现的”或超出范围的值，且易于保存原分布的形状。其缺点为所有的预测变量必须是分类别的（或被当做如此），因此对可能预测变量的数目进行了严格的限制。为了去除这个限制，Little(1988)提出了一个叫做预测均数匹配的部分参数方法。如同多变量正态参数方法一样，这个方法开始也基于有完整数据的个案，用 Y 对一组变量做回归，再由得到的回归模型产生有缺失和没缺失数据个案的预测值。接着，对于有缺失数据的每一个个案，我们找一组有完整数据的个案，其 Y 的预测值与有缺失数据个案的 Y 预测值是“接近”的。从这组个案，我们随机选择一个个案的 Y 值捐赠给该缺失个案。

对于单一变量 Y，可以直接将“接近”定义为预测值间的绝对差异。然而，接下来我们必须决定多少个接近的预测值应被包含进每一个缺失个案的捐赠库中？或相同地，构成可能的捐赠值组的接近的截略点应为何？如果选取了较小的捐赠库，在估计值中就会有较多的抽样变异性。另一方面，太大的捐赠库则容易导致可能的偏误，因为许多捐赠者可能与受援者不相像。为了处理这个模棱两可的情况，Schenker

和 Taylor(1996)提出了一个“适合的方法”，根据接近预测值的完整个案的“密度”，变化每一个缺失数据个案捐赠库的大小。他们发现，该方法比其他有固定的3个或10个捐赠库的方法稍微好些。然而，这三个方法间的差异太小以至于该方法几乎不值得我们去花额外的计算成本。

执行预测均数匹配时，因为回归系数只是真实系数的估计值，因此做些调整也是必要的。和参数的例子一样，这可以通过在计算每一个插补数据组的预测值前，先从它们的后验分布中随机抽取一组新的回归参数来完成。以下给出具体的操作步骤：

(1) 对于没有缺失 Y 数据的 n_1 个个案回归 Y 于 X(一协变量的向量)，产生回归系数 b(一个 $k\times 1$ 向量)及残差方差估计值 s^2；

(2) 从一个(假设不提供信息的先验的)残差方差的后验分布中进行随机抽样。可通过计算 $(n_1-k)s^2/\chi^2$ 来完成，其中 χ^2 代表从一个有 n_1-k 个自由度的卡方分布中随机抽取而得。令 $s_{[1]}^2$ 为第一个这样的随机抽取；

(3) 从回归系数的后验分布中随机抽取系数值。可通过从一个有着平均数为 b 及协方差矩阵 $s_{[1]}^2(X'X)^{-1}$ 的多变量正态分布抽取所完成，其中 X 为一个 X 值的 n_1-k 矩阵。此外，$b_{[1]}$ 为第一个这样的随机抽取。

有关如何执行的实用建议参见 Schafer(1997)。

对于每一组新的回归参数，所有个案产生预测值。接着对于每一个 Y 有缺失数据的个案，我们根据预测值建构一个捐赠者库，并随机从该捐赠库中选择一个 Y 的观察值。虽然计算可能会变得相当复杂，但这个预测均数匹配法可以被广

义化至超过一个 Y 变量有缺失数据的情况(Little, 1998)。

经验残差抽样法

在数据扩增法中，残差值从一个标准正态分布抽样而来，并且残差值接着被加到预测回归值中，以得到最终的插补值。我们可以通过在线性回归所产生的实际残差组中做随机抽取，来修正这个方法，从而使它较少依靠参数假设。这可以产生插补值，其分布更像该被观察到的变量的分布(Rubin, 1987)，虽然这个方法仍有可能得到落于允许区间外的插补值。

如同用其他方法进行多重插补一样，存在与正确执行这个方法涉及许多相关的重要的细节问题。和之前一样，令 Y 为有缺失数据的变量，用 n_1 个有观察数据的个案来插补 n_0 个个案。令 X 为没有缺失数据的 n_1 个个案之 $k\times 1$ 变量向量(包含一个常数项)。我们由执行前述的三个步骤开始，以获得 Y 对于 X 的线性回归，并且从参数的后验分布中产生随机抽取值。接着我们增加下列步骤：

(4) 根据步骤 1 中的回归估计值，对有缺失数据的个案计算标准化的残差：

$$e_i = (y_i - bx_i)/\sqrt{s^2(1-k/n_1)}$$

(5) 从步骤 4 计算得到的 n_1 个残差中抽取一个有 n_0 个值的简单随机样本(有替换)。

(6) 对于有缺失数据的 n_0 个个案，计算 Y 的插补值，如：

$$y_i = b_{[1]}x_i + s_{[1]}e_i$$

其中 e_i 代表步骤 5 中抽取的残差,而 $b_{[1]}$ 和 $s_{[1]}$ 是从参数的后验分布的第一次随机抽取得到的值。

这 6 个步骤产生一个完整的数据组。为了得到额外的数据组,简单地重复步骤 2 到 6 即可(除了不应该被重复的步骤 4 以外)。

Rubin(1987)曾经解释过,这个方法可以很容易被延伸至数个变量上有着单调缺失形态的数据组。当其为缺失时,每一个变量使用所有被观察到的变量作为预测量来进行插补。这个经验残差方法也可以被修正以考虑插补值中的异方差性(Schenker & Taylor, 1996)。对于每一个插补的个案,残差库被限制为那些 Y 预测值与缺失数据个案的 Y 预测值接近的被观察到的个案。

实例

让我们对大专院校数据的一个次集合试试看部分参数方法。对 1272 所大专院校,TUITION 是全部都被观察到的(为了简便,我们应该去掉在该变量上有缺失数据的 30 个个案)。在这 1272 所大专院校中,只有 796 个报告 BOARD,即各大专院校的年度平均食宿费用。使用 TUITION 作为一个预测量,我们的目的是要对其他 476 所大专院校插补 BOARD 的缺失值,并且对所有 1272 所大专院校估计 BOARD 的平均数。

首先,我们运用先前使用过的方法。对于有完整数据成列删除的 796 所大专院校,BOARD 的平均值为 2060 美元,其标准误为 23.4。对 TUITION 和 BOARD 运用 EM 运算

法，我们得到 BOARD 的平均数为 2032（但没有标准误）。BOARD 和 TUITION 间的 EM 相关性估计值为 0.555。使用数据扩增法、多变量正态模型下的多重插补产生的 BOARD 的平均数为 2040，估计标准误为 21.2。

因为 BOARD 高度右偏，故有理由怀疑多变量正态模型可能不是很恰当。相当多由数据扩增法所插补的值比最低的被观察值 531 小，且有一个插补值是负的。或许我们可以通过在经验残差上抽样从而做得更好。对于有 TUITION 和 BOARD 两者数据的 796 个个案，回归 BOARD 于 TUTION 的普通最小二乘（OLS）回归方程为：

$$BOARD = 1497.4 + 67.65 \times TUITION/1000$$

其均方根差估计为 542.6。这个回归方程计算得到的标准化残差由 796 个个案计算所得。

估计得到的回归参数被用于从参数的后验分布中做五次随机抽取，如同在步骤 2 和步骤 3 中（假设一个不提供信息的先验分布）。被抽取的值为：

截　距	斜　率	均方根差
1536.40	66.6509	531.900
1503.65	71.5916	552.708
1501.61	66.9756	554.800
1486.84	66.9850	548.400
1504.23	61.2308	534.895

为了创造第一个完整数据组，476 个残差值从 796 个个案中被随机有替代的抽取而产生。这些标准化残差被任意地指派给 BOARD 有缺失数据的 476 个个案。令 E 为某一给定个案的指派残差，BOARD 的插补值产生如下：

$$\text{BOARD} = 1536.40 + 66.6509 \times \text{TUITION}/1000 + 531.90 \times E$$

对四个剩余的数据组重复这个过程，并于每一个步骤中，在残差上有新的抽样，回归参数有新的值。

一旦产生五个数据组，就对每一个数据组计算其平均数和标准误，并使用标准误的方程 5.1 来计算结果。BOARD 的最终平均数估计值为 2035，估计标准误为 20.4，相当接近基于正态分布的多重插补。

现在，让我们尝试预测均数匹配。根据从 BOARD 对于 TUTION 的 OLS 回归的系数，作者从参数的后验分布中进行五次新的随机抽取：

截　距	斜　率	均方根差
1465.89	67.8732	557.531
1548.98	64.5723	539.952
1428.82	67.3901	512.381
1469.34	67.3750	550.945
1517.92	66.1926	534.804

对于第一组参数值，作者对所有观察到的及缺失的个案推测 BOARD 的预测值。对于每一个 BOARD 有缺失数据的个案，作者找到了预测值最接近缺失个案预测值的五个个案。同时随机选择这五个个案中的其中一个，并指派它被观察到的 BOARD 值作为插补值给缺失的个案。对五组参数值的每一组都重复这个过程，以产生五个完整的数据组（数据组数目与被观察到的个案数目匹配到每一个缺失个案的数目相同只是巧合罢了）。接着对每一个数据组计算平均数和标准误并用通常的方法结合结果。BOARD 的结合平均数

为 2028,其估计标准误为 23.0。

所有四个插补方法都产生相似的平均数估计值,且所有都显著低于根据成列删除而得的平均数。Schenker 和 Taylor(1996)建议,虽然参数和部分参数插补方法倾向产生相似的平均数结构(包含回归系数)的估计值,但它们可能对于插补值的边际分布产生更加不同的结果。他们的模拟研究显示,对于主要关注点是边际分布的应用,部分参数模型有一个明显的优点。当被用来产生估计值的回归以许多方式被错误设定时,更是如此。

第7节 | 连续的广义回归模型

数据扩增法的吸引力之一是,不像其他刚讨论过的参数或半参数方法,它可以很容易地处理有缺失数据的、有大量变量的数据组。遗憾的是,这个方法需要对所有变量指明一个多变量分布,而当变量有着多种类型的时候(例如,连续的、二元的及计数数据),这并不是件容易的事。有另一种方法被提议用来处理有着数种不同变量类型的大的复杂数据组之缺失数据。不拟合一单一综合模型(如多变量正态),而是对每一个有缺失数据变量分别指明一个回归模型。这个方法涉及在数个回归模型中循环,在每一个步骤插补缺失值。

虽然这个方法非常吸引人,但它却不像我们已讨论过的其他方法那样具有很强的理论上的有效性。在本书写作之时,详细的论述只有 Brand(1999)、Van Buuren 和 Oudshoorn 以及 Raghunathan、Lepkowski、Van Hoewyk 和 Solenberger (1999)等人未发表的报告。在 Raghunathan 等人关于此方法的论述中,可利用的模型包含正态线性模型、二元 logistic 回归、多项 logit 回归及 poisson 回归。回归模型以一个特定顺序来估计,从有着最少缺失值的因变量进行到有最多缺失数据的因变量。我们用 Y_1 至 Y_k 来表示这些变量,并令 X 表

示没有缺失数据的该组变量。

第一个"回合"的估计进行如下。回归 Y_1 于 X，并用相似于"对于多重插补的其他参数的方法"部分已叙述过的、对多项的 logit 模型产生插补值的方法以产生插补值。可以对插补值设定边界和限制。接着，回归 Y_2 于 X 和 Y_1，包含 Y_1 的插补值，并且产生 Y_2 的插补值。然后，回归 Y_3 于 X、Y_1 和 Y_2（含两个 Y 的插补值），继续直到所有回归都被估计过了。第二个和后续的回合重复这个过程，但现在每一个变量要对所有使用从前面步骤产生的插补值的其他变量做回归。这个过程继续至某一个预先指定的回合数目或直到出现稳定的插补值。一个可以完成这些任务的 SAS 的宏命令可参见 http://www. isr. umich. edu/src/smp/ive。

Van Buuren 和 Oudshoorm 把他们的这个方法命名为基于链式方程的多重插补，而且他们开发了 S-PLUS 程序以执行它（参见 http://www. multiple-impuation. com/）。他们的方法和 Raghunathan 等人的方法的主要差异在于不包含泊松分布，但对于插补值的随机抽取方法却允许有更多的选择（包括参数和部分参数）。

第8节｜线性假设检验和最大似然比检验

我们所使用的多重插补的统计推论方法一直都非常简单。对于一个给定参数,该估计值的标准误通过方程5.1来计算。这个标准误接着被插入立足于正态近似反映的传统方程中,以对某些关注的假设产生一个置信区间或一个t统计量。有时候,这样并不足够。通常我们想要对参数组检验假设,例如,两个参数相等或数个参数全部都等于0。当我们对一组虚拟变量估计数个系数时,这些种类的假设显得特别重要。此外,有必要通过比较一个模型与另一个较简单的模型来计算似然比统计量。当执行多重插补时,完成这些任务并不是那么容易。Schafer(1997)叙述过三种不同的方法,但没有一种是完全令人满意的。作者在此简短地叙述这三种方法,同时也会在下一个部分给出一个实例。

使用结合协方差矩阵的Wald检验

当没有缺失数据时,一个关于多元参数推论的普遍方法为,根据参数估计值及其估计的协方差矩阵来计算Wald卡方统计量。这里有简单的回顾,但遗憾的是,需要矩阵代数

学。假设我们想要估计一个 $p\times1$ 参数向量 β。我们有估计值 $\hat{\beta}$ 和估计的协方差矩阵 C,我们想要检验一个表达为 $\mathrm{L}\beta=\mathrm{c}$ 的线性假设,其中 L 为一个 $r\times p$ 的常数项矩阵,且 c 为一个 $r\times1$ 常数项向量。例如,如果我们想要检验 β 的前两个元素彼此相等这个假设,我们需要 $\mathrm{L}=[1\ -1\ 0\ 0\ 0\cdots0]$ 及 $\mathrm{c}=0$。Wald 检验的计算如下:

$$W=(\mathrm{L}\hat{\beta}-\mathrm{c})'[\mathrm{LCL}'](\mathrm{L}\hat{\beta}-\mathrm{c}), \qquad [6.1]$$

其在零假设下,有一个自由度为 r 的近似反映卡方分布。[13]

现在我们广义化这个方法到多重插补的情形中。不用 $\hat{\beta}$,我们可以使用 $\bar{\beta}$,即横跨数个完整数据组的估计值之平均数,也就是:

$$\bar{\beta}=\frac{1}{M-1}\sum_{k}\hat{\beta}_k$$

其次我们需要一个结合样本内变异与样本间变异的协变量矩阵的估计值。令 C_k 为在数据组 k 中的参数估计的协方差矩阵,且令 $\bar{\mathrm{C}}$ 为那些横跨 M 个数据组矩阵的平均。样本间变异被定义为:

$$\mathrm{B}=\frac{1}{M-1}\sum_{k}(\hat{\beta}_k-\bar{\beta})(\hat{\beta}_k-\bar{\beta})'$$

协方差矩阵的结合估计值则为:

$$\tilde{\mathrm{C}}=\bar{\mathrm{C}}+(1+1/M)\mathrm{B}$$

这只是方程 5.1 的没有平方根的一个多变量广义化。我们用包含 $\bar{\beta}$ 和 $\tilde{\mathrm{C}}$ 的方程 6.1,来替代 $\hat{\beta}$ 和 C,以得到我们的检验统计量。

遗憾的是,在典型的 M 小于或等于 5 的例子中,这起不

到好的作用。在诸如此类的例子中，B是一个对于协方差矩阵相当不稳定的估计量，且造成的 W 的分布不是卡方。Schafer(1997)给予一个对协方差矩阵更加稳定的估计量，但这需要不合理的假设，即假设对于 $\hat{\beta}$ 的所有元素，缺失信息比都相同。然而，有些模拟研究显示，即使当假设被违反时，这个备选方法仍行得通。这个方法已被纳入 SAS 程序 MIANALYZE 中。

似然比检验

如果感兴趣的模型通过最大似然来估计且没有缺失数据，通常通过计算似然比卡方检验来执行多参数检验。这个程序相当简单。令 l_0 为强加假设时模型的对数似然，并令 l_1 为放松假设时模型的对数似然。似然比统计量即为 $L = 2(l_1 - l_0)$。

与之前一样，我们的目的是要将这个广义化到多重插补中。第一个步骤是要对 M 个完整数据组中的每一个执行想要的似然比检验。令 $\bar{L}$ 为横跨 M 个数据组的似然比卡方的平均数，这是较容易的部分。接下来是比较困难的部分。为了得到那些卡方，必须要在每一个数据组中估计两个模型，即有强加假设的模型和放松假设的模型。令 $\bar{\beta}_0$ 为强加假设时，M 个参数估计值的平均数，且令 $\bar{\beta}_1$ 为放松假设时，M 个参数估计值的平均数。在每一个数据组中，我们接着计算一个参数值被限制为 $\bar{\beta}_0$ 的模型的对数似然，且再次计算一个参数值被设为 $\bar{\beta}_1$ 的模型的对数似然（这明显需要能够计算和报告使用者指定参数值的对数似然比的软件）。根据这两

个对数似然，在每一个数据组中计算一个似然比卡方。令 $\tilde{L}$ 为这些横跨 M 个样本的卡方统计量的平均数。

最终的检验统计量则为 $\tilde{L}/\left(r+\left(\frac{M+1}{M-1}\right)(\bar{L}-\tilde{L})\right)$，其中 r 为假设所强加的限制数目。在零假设下，这个统计量有着近似反映分子自由度为 r 的 F 分布。分母自由度的计算有点困难。令 $t=r(M-1)$ 且令

$$q=\left(\frac{M+1}{M-1}\right)\left(\frac{\bar{L}-\tilde{L}}{r}\right)$$

如果 $t>4$，则 d. d. f. $=4+(t-4)[1+(1-2/t)/q]^2$。如果 $t\leqslant 4$，则 d. d. f. $=t(1+1/r)(1+1/q)^2/2$。

结合卡方统计量

Wald 检验和似然比检验，两者都缺乏前面使用的单一参数方法的简单性。而且它们需要有专业选项及输出的分析软件，这是我们通常都尽量避免的。

作者现在讨论一个容易从标准输出所计算的第三个方法，但该方法可能不像另外两种方法那样精准(Li, Meng, Raghunathan & Rubin, 1991)。该方法所需要的仅是计算在 M 个完整数据组中，每一个数据组的传统的卡方统计量(Wald 或似然比)，以及相关的自由度。

令 d_k^2 为数据组 k 中有着自由度为 r 的卡方统计量，令 $\bar{d}^2$ 为 M 个数据组的这些统计量的平均数，且令 s_d^2 为 M 个数据组的卡方统计量的平方根的样本方差，也就是：

$$s_d^2=\frac{1}{M-1}\sum_k(d_k-\bar{d})^2$$

被提议的检验统计量为:

$$D=\frac{\bar{d}^2/r-(1-1/M-1)s_d^2}{1+(1+1/M-1)s_d^2}$$

在零假设下,这个统计量有着近似反映分子自由度为 r 的 F 分布。分母自由度近似反映:

$$\left(\frac{M-1}{r^{3/M}}\right)\left(1+\frac{M}{(M+1/M)s_d^2}\right)^2$$

作者已经写了一个 SAS 的宏命令(COMBCHI)以执行这些运算和计算一个 p 值。该命令可以参见作者的网站(http://www.ssc.upenn.edu/~Allison)。你只需要键入数个卡方值以及自由度,宏命令就会报告一个 p 值。

第 9 节 | MI 实例 2

让我们考虑另一个详细的经验实例来说明本章所讨论过的某些技巧。数据组来自 1994 年综合社会调查，共有 2992 名受访者(Davis & Smith, 1997)。我们的因变量为 SPANKING，一个对于"有时候管教一个小孩责打是必要的。你强烈赞成、赞成、反对或强烈反对"这个问题的回答。如同问题本身所提示的，有四个可能的依序的答案，被编码为 1 至 4 的整数。这个问题被设计作为只对随机的 2/3 的样本施行的模块的一部分。因此，有 1015 个完全随机缺失的个案。此外，另有 27 个缺失个案，其回答被编码为"不知道"或"没有回答"。

我们的目的是要估计一个依序的 logistic(累进的 logit)模型(McCullagh, 1980)，其中通过下列变量预测 SPANKING：

AGE 受访者的年龄以岁计算，从 18 到 89。缺失 6 个个案。

EDU 受教育年数的数目。缺失 7 个个案。

INCOME 家庭收入，以 21 个区间类别的终点编码，以千元计算。缺失 356 个个案。

FEMALE 1 = 女性；0 = 男性。

BLACK 1 = 黑人；0 = 白人，其他。

MARITAL 婚姻状态的 5 个类别。缺失 1 个个案。

REGION 9 个区域类别。

NOCHILD 1 = 没有小孩;其他为 0。缺失 9 个个案。

一个额外的变量 NODOUBT 需要更进一步的解释。受访者被问到他们对上帝的信仰,有 6 个回答类别,从“我不信上帝”到“我知道上帝真的存在而且我对于这点没有疑惑”。有 62%的受访者的典型回答是“我知道上帝真的存在而且我对于这点没有疑惑”。然而,如同责打那个问题一样,这个问题只是问卷模块的一部分,只随机询问 1386 个受访者中的一部分。因此,根据设计有 1606 个缺失个案。其余 60 个缺失是因为他们说“不知道”或“没有回答”。如同这里所使用的,如果受访者“没有疑惑”变量被编码为 1,反之则会被编码为 0。

大多数的缺失数据在 3 个变量以上:SPANKING、NODOUBT 和 INCOME。样本中有 5 个主要的缺失形态,占了 96%的受访者:

771 个个案　在任何变量上都没有缺失;

927 个个案　只有 NODOUBT 缺失;

421 个个案　只有 SPANKING 缺失;

89 个个案　只有 SPANKING 缺失;

509 个个案　SPANKING 和 NODOUBT 缺失;

421 个个案　NODOUBT 和 INCOME 缺失。

如往常一样,数据分析最简单的方法是成列删除,其只使用 26%的原始样本。为了指明模型,作者对婚姻状态类别创造虚拟变量:NEVER(从未结婚)、DIVSEP(离婚或分居)及 WIDOW(鳏寡),以已婚者为参照组。对区域也生成三个虚拟变量(以西方作为省略的类别)。[14](由 SAS 中 PROC LOGISTIC 所产生的)结果显示于表 6.3 的第一栏中。黑人、

表 6.3 预测 SPANKING 累进的 logit 模型的系数估计值(及标准误)

变　量	成列删除	正态数据扩增法	连续回归	连续回归（缺失数目于SPANKING 上）
FEMALE	−0.355(0.141)*	−0.481(0.089)***	−0.449(0.098)***	−0.489(0.094)***
BLACK	0.565(0.218)**	0.756(0.117)***	0.693(0.119)***	0.685(0.135)***
INCOME	−0.0036(0.0033)	−0.0052(0.0020)*	−0.0042(0.0027)	−0.0047(0.0022)*
EDUC	−0.0055(0.027)*	−0.061(0.016)***	−0.073(0.019)**	−0.068(0.016)***
NODOUBT	0.454(0.147)**	0.465(0.120)**	0.455(0.156)*	0.438(0.121)**
NOCHILD	−0.205(0.199)	−0.109(0.112)	−0.141(0.164)	−0.091(0.123)
AGE	0.010(0.005)*	0.0043(0.0032)	0.0031(0.0032)	0.0040(0.0031)
EAST	−0.712(0.219)**	−0.444(0.125)***	−0.519(0.156)**	−0.488(0.136)***
MIDWEST	−0.122(0.203)	−0.161(0.136)	−0.228(0.149)	−0.159(0.128)
SOUTH	0.404(0.191)**	0.323(0.156)*	0.262(0.129)*	0.357(0.121)**
NEVMAR	−0.046(0.238)	−0.075(0.148)	−0.036(0.173)	−0.071(0.151)
DIVSEP	−0.191(0.194)	−0.203(0.150)	−0.141(0.128)	−0.184(0.126)
WTDOW	0.148(0.298)	−0.244(0.150)	−0.116(0.177)	−0.215(0.174)

注：* $p < 0.005$，** $p < 0.01$，*** $p < 0.001$。

较老的受访者及对于上帝"没有疑惑"的受访者有可能赞同责打小孩有时是必要的。女性及受更多教育的受访者则可能比较反对。也有大的区域性差异:来自南方的受访者对于责打更加赞同,而来自东北的则更加反对。另一方面,没有证据显示收入、婚姻状态或有小孩等有任何效应。

因为依据设计有84%观察值有缺失数据(且因此为完全随机的),成列删除应该产生近似反映无偏误的估计值。然而,损失几乎3/4的样本是要付出重大代价的,而这个代价若使用多重插补则是可以避免的。为了执行MI,作者首先使用第5章叙述过的多变量正态模型下的数据扩增法。执行之前,先删除婚姻状态有缺失值的个案,以避免必须插补一个多类别变量。对去掉缺失数据的SPANKING变量的1042个个案,有一个合理的论点可以被提出,因为因变量上有缺失的个案包含很少关于回归系数的信息。然而,包含它们并没有损失而且可能还会有一些好处,所以作者把他们全部保留了。在模型中所有13个变量被全部纳入插补过程,没有经过任何正态化的转换。

所有虚拟变量的插补值被四舍五入至0或1。SPANKING的插补值被四舍五入至1到4的整数。年龄和收入有一些正当范围外的插补值,而这些值被重新编码至上限或下限。累进的logit模型接着被用来对五个数据组通过标准公式进行估计,得到的估计值再合并在一起。

结果显示在表6.3的第2栏中。结果的基本模式是相同的,这一栏中INCOME的效应是显著的,但AGE变得不显著了。最引人注目的是,所有系数的标准误都比那些成列删除的标准误低很多,最典型的低了约40%。甚至

NODOUBT 的标准误也小了 18%，这是非常惊人的，因为超过一半的个案在该变量有缺失。对于许多变量较小的标准误产生低很多的 p 值。

累进的 logit 模型强加一个被称为成比例发生比假设以限制数据。简单地说，这个词组指对于任何因变量，其二分化系数被假设为相同的。PROC LOGISTIC 对成比例发生比假设为正确的这个零假设，报告一个卡方统计量（分数检验）。但由于我们处理五个数据组，因而得到五个卡方：32.0、31.3、38.0、36.4 和 35.2，每一个自由度都为 26。使用前述宏命令 COMBCHI，这五个值被结合以产生一个 0.25 的 p 值，暗示模型强加的限制能很好地拟和数据。对于这五个数据组中的每一个，作者也计算了所有区域系数都为 0 的零假设下的 Wald 卡方。自由度为 3，这些 Wald 卡方值为 72.9、81.3、53.4、67.7 和 67.0。结合的 p 值为 0.00002。

在表 6.3 的第 3 栏中，我们看到运用 Raghuanthan 等人（1999）的多重插补方法的结果，这些结果依靠连续的广义模型。对于每一个有缺失数据的变量估计一个回归模型，将该变量当做因变量，其他所有变量当做预测量。这些回归模型接着被用来产生五组随机插补值。对 EDUC 和 INCOME 而言，虽然上下限被嵌入插补过程中，但该模型为普通线性模型。logistic 模型则指明给 NODOUBT 和 NOCHILD。一个多项 logit 模型被用于 SPANKING。在插补过程中有 20 个回合，也就是说，对于五个完整数据组中的其中任意一个，直到得出最终结果前，有缺失数据的变量会被连续地插补 20 次。

同之前一样，累进的 logit 模型被用来估计五个完整数据

组中的每一个,并使用公式结合结果。表6.3第3栏中的系数估计值与那些用多变量正态数据扩增法得到的系数估计值相当类似。标准误通常会比数据扩增法的标准误大,虽然不像成列删除的标准误那么高。

令人意外的是,连续回归的成比例发生比假设卡方统计量几乎是正态数据扩增法的成比例发生比假设的卡方统计量的两倍大。更明确的是,自由度为26,其值为54.9、59.9、66.7、85.4和59.0,每一个都有一个小于0.001的p值。然而,当用宏命令COMBCHI结合这些值时,得到的p值为0.45。为什么个别的p值和结合p值间会有这么大的差异呢?其答案在于,在卡方间的大的方差表示它们中的每一个可能都是严重高估的值。结合它们的方程已经考虑到这点了。

那么正态扩增法和连续回归间卡方的不同又如何解释呢?作者怀疑这源自插补SAPNKING的多项的logit模型没有对该变量强加任何顺序的事实。因此,插补值不可能与成比例发生比假设相符。当以一个SPANKING的线性模型重做连续插补时(将插补值四舍五入到整数),成比例发生比假设的卡方与那些在正态数据扩增法下获得的卡方更加一致。二者择其一,作者在先删除所有SPANKING缺失个案后重做连续插补。SPANKING仍然被指名为类别的,亦即当插补其他变量的值时,它被视为一个类别的预测量。而且,成比例发生比假设的卡方跟那些由正态数据扩增法产生的卡方相似。

表6.3的最后一栏显示删除SPANKING缺失个案后连续回归插补的结合结果。有趣的是,两者系数和它们的标准误普遍地更接近于数据扩增法。更进一步讲,当我们删除1042个SPANKING有缺失数据的个案时,并没有明显的信息损失。

第 10 节｜长期的及其他集群数据的 MI

到目前为止，我们已假设每一个观察值独立于其他观察值，如果数据为每个大总体中的简单随机样本，这即为一个合理的假设。然而，许多数据组可能在各观察值间有某种相关性。例如，假设我们有一个面版个体数据，对他们而言，连续五年每年都测量相同的变量。许多分析面版数据的计算机程序需要组织数据，将每一年的测量当做个别的观察值。为了将观察值连结在一起，必须有一个变量包含识别号码，而识别号码对于来自相同个体的所有观察值都是一样的。

因此，如果我们有 100 个连续五年被观察的个体，我们会有 500 个有效的观察值。显而易见，这些观察值不会是独立的。如果已讨论过的多重插补方法被直接运用于这 500 个观察值上，不会利用到任何长期的信息。因此，完整数据组可能会产生长期相关性的、严重的低估值，特别是如果有大量的缺失数据的话。

如果观察值自然落入自然发生的集群中，也会产生相似的问题。假设我们有一个 500 对已婚者的样本，且对夫妻二人都询问相同的问题。如果我们对配偶中任选其一插补缺失数据，那么使用配偶间答案的相关性是很重要的。相同的也适用于在同一教室内的学生或同一小区内的受访者。

一个对于这些数据类别的处理方法是,在一个嵌入观察值间相关性的模型下进行多重插补。Schafer(1997)对集群的数据提出了一个多变量线性混合效应模型,并且也开发了一个计算机程序(PAN)以使用数据扩增法执行插补(参见http://www.stat.psu.edu/~jls/)。虽然之前承诺会有一个Windows版本,但目前仅以S-PLUS套装软件的一个程序库来运行。

有一个更简单的方法,当调查的次数相对较少时,对于跟踪数据可以处理得很好。基本的想法是要格式化数据,以致对于每一个个体只有一个记录,而且在不同时点对同一变量有不同的测量。多重插补接着以我们所考虑过的任何一种方法来执行。这就把在任何时点的变量,用来作为任何其他时点的变量的预测量的情况考虑进去了。一旦数据被插补,数据组就可以被重新格式化,以致对于每一个个体都有数笔记录,其对每一个时点都有一笔记录。

第 11 节 | MI 实例 3

这里有一个比较简单的使用刚讨论过的方法进行长期数据多重插补的例子。样本由 220 名白人女性所组成，年龄至少都在 60 岁以上，在大宾州地区进行髋关节骨折手术(Mossey, Knott & Craik, 1990)。在她们出院后，她们被访问三次：2 个月后、6 个月后及 12 个月后。下列五个变量在三次访问中每一次都会被测量。

CESD 一个忧郁程度的测量，范围从 0 至 60。

SRH 自我健康评估，以一个包含 4 个值的量表测量(1=不良，4=杰出的)。

WALK 如果病患在家不用帮助可以行走，编码为 1；反之为 0。

ADL 可以不用协助而完成的自理“日常生活活动”的数目。

PAIN 病患所经历的疼痛程度范围，从 0(没有)到 6(持续)。

我们的目的是要以 CESD 作为因变量及其他四个作为自变量来估计一个“固定效应”线性回归模型(Greene, 2000)。模型的形式为：

$$y_{it} = \alpha_{it} + \beta_1 x_{it1} + \cdots + \beta_4 x_{it4} + \varepsilon_{it}$$

其中 y_{it} 为个人 i 在时点 t 的 CESD 值,而 ε_{it} 满足线性模型的通常假设。这个模型值得注意的是,对于样本中每一个个人,都有一个不同的截距 α_i,从而控制病人们所有稳定的特性。这个个人特定的截距也引入了每一个个人的多个回答间的相关性。

为了估计模型,还创造了一个有 660 个观察值的可用的数据组,对每一个时点每一个个人都有一笔观察值。要得到 OLS 回归估计值有两个相同的计算方法:(1)对于每一个个人(小于 1)包含一个虚拟变量;(2)执行回归于离差分数上。第二个方法在执行多元回归前,先在模型中用每一个变量减去(跨三个时点的)个人特定的平均数。

遗憾的是,该研究有大量的流失,以及各个时点上额外的无回答。如果我们删除所有具任何缺失数据的个人时点,可用的数据组会从 660 个个案减少至 453 个个案。如果我们删除在任何时间任何变量有缺失数据的个人,则数据组会减少为 101 个个人(或 303 个个人时点)。

表 6.4 显示使用四种方法处理缺失数据的固定效应回归结果。[15]前两栏给出了从两种成列删除版本得来的系数和标准误:(1)删除有任何缺失数据的个人;(2)删除有任何缺失数据的个人时点。有一个明显的证据显示,忧郁程度被自我健康评估所影响,对于行走能力的影响则不足为凭。忧郁程度在第 1 次和第 2 次调查中明显地比第 3 次(当大部分的病患已完全痊愈)高出许多。很少或没有证据显示 ADL 和 PAIN 有效应。

最后两栏给出了根据全部样本,含已在多变量正态模型下以数据扩增法插补的缺失数据的分析结果。[16]第 3 栏的

表 6.4 预测 CESD 的固定效应模型之系数估计值(及标准误)

	成列删除，以个人	成列删除，以个人时点	数据扩增法，以个人时点	数据扩增法，以个人
SRH	2.341(0.586)**	1.641(0.556)**	2.522(0.617)**	1.538(0.501)**
WALK	−1.552(0.771)*	−1.381(0.761)	−1.842(0.960)	−0.550(0.825)
ADL	−0.676(0.528)	−0.335(0.539)	−0.385(0.562)	−0.410(0.435)
PAIN	0.031(0.179)	0.215(0.168)	0.305(0.180)	0.170(0.164)
WAVE 1	8.004(0.650)**	8.787(0.613)**	6.900(0.729)**	9.112(0.615)**
WAVE 2	7.045(0.579)**	7.930(0.520)**	5.808(0.642)**	8.131(0.549)**
N(个人时点)	303	453	660	660

注：* $p < 0.05$，** $p < 0.01$。

结果,对被视为是独立观察值的660个人时点进行插补。因此,缺失数据仅通过同一时点的信息来插补。为了进行最后一栏的插补,数据被重新组织为220个个人,在每一个时点有区分的变量名称。这样一来,每一个有缺失数据的变量根据所有三个时点的信息以进行插补。原则上,这应该产生比较好的插补,特别是因为一个缺失值可以被不同时点的同一变量的测量值所预测。

事实上,最后一栏所有的估计标准误都比倒数第2栏的小一些。它们也比两种成列删除方法中的任一种更小一些。另一方面,根据个人次数时点的数据扩增法之标准误则比两种成列删除方法稍微大些。无论如何,在这个应用中,多重插补并不具有压倒性的优势。从质量上讲,不管使用何种插补方法,结论几乎都是相同的。

第7章

不可忽略的缺失数据

前面几章着重讲述了在缺失数据机制可以忽略的情况下能够使用的方法。可忽略性意味着我们不需要对数据发生缺失的过程进行模型化。可忽略性重点要求数据是随机缺失的——特定变量缺失数据的概率不取决于该变量的值(扣除分析中其他变量的作用之后)。

处理可忽略的缺失数据的基本策略可以简单地总结如下:调整所有缺失和非缺失数据间的可观察到的差异,并且假设所有剩余的差异为无系统性的。当然,这是一个熟悉的策略。标准回归模型就被设计来做这个——调整可观察到的差异并假设所有未观察到的差异为无系统性的。

遗憾的是,通常我们有足够的理由怀疑数据不是随机缺失的。例如,常识告诉我们,曾被逮捕过的人比不曾被逮捕过的人更不可能报告他们的逮捕状态,有高收入的人可能不会报告他们的收入。在临床药物试验中,变得更差的人比那些变得更好的人,更有可能退出临床药物试验。

在这些情况下应该怎么办呢?有处理不可忽略的缺失数据的模型且想要运用它是很自然的。然而,意料之中的是,很少有可用的软件可以估计不可忽略的模型(除了一个重要的例外——Heckman 的选择性误差模型)。其基本问题

在于，对数据给定一个模型，只有唯一一个可忽略缺失数据机制，但有无限多不同的不可忽略的缺失数据机制。所以很难编写计算机程序，以处理即使是这些可能性的一小部分。此外，根据选择的模型，答案可能会变化很大。所以选择正确的模型非常重要，而该选择需要对所调查的现象有非常精准及详尽的知识。更糟糕的是，没有经验方法可以从一个不可忽略的模型（或从一个可忽略的模型）中分别出另一个“不可忽略的模型”。

也许你不会极端到说：“不要做那个”，但你可能会这么说：“如果你做那个，要特别小心。”此外，如果你没有很多统计方面的专业知识，那么请找一个具备统计专业知识的合作者。本章针对处理不可忽略的缺失数据的一些方法，为大家提供一个简短的导论及概要。

你所需要明确的第一件事情是，作者已经介绍过的对于可忽略的缺失数据的方法——最大似然和多重插补——可即刻适用于处理不可忽略的缺失数据。假设选择的模型正确，那么这两个方法有如他们在可忽略的设定下的相同的最适特性。第二点要记住的是，任何有关不可忽略的缺失数据的方法应该要伴随一个敏感性分析。因为根据假设的模型，结果可能变化很大，故试验一貌似有理范围的模型并看它们是否产生相同的结果是很重要的。

第1节 | 两种模型

不管你选择最大似然还是多重插补，处理不可忽略的缺失数据有两种截然不同的方法：选择模型和形态混合模型。对一个缺失数据的单一变量解释最为容易。令 Y 为关注的变量且令 R 为一个虚拟变量，如果 Y 被观察到，其值为 1，如果 Y 缺失，则其值为 0。令 $f(Y, R)$ 为这两个变量的联合概率密度函数。选择一个模型意味着对 $f(Y, R)$ 选择某些明确的指定。

可以使用两种不同的方法因子化联合 p. d. f(Little & Rubin, 1987)。在选择模型中我们使用：

$$f(Y, R) = \Pr(R \mid Y) f(Y)$$

其中 $f(Y)$ 为 Y 的边际密度，且 $\Pr(R|Y)$ 为给定某 Y 值下的 R 的条件式概率。用文字表示，如同没有缺失数据一样我们首先对 Y 建立模型。例如，我们可以假设 $f(Y)$ 为有着平均数 μ 和方差 σ^2 的正态分布，且 $\Pr(R|Y)$ 为：

$$\Pr(R = 1 \mid Y) = \begin{cases} p_1 & if\ Y > 0 \\ p_2 & if\ Y \leqslant 0 \end{cases}$$

这个模型被确认且可以被 ML 所估计。对应形态混合模型的备选的联合 p. d. f 因子化公式为：

$$f(Y,R) = f(Y \mid R)\Pr(R)$$

其中 $f(Y, R)$是条件为是否缺失的 Y 的密度。例如，我们可以假设 $\Pr(R)$就是某个固定的 θ，且是有着方差为 σ^2 和当 $R=1$ 时平均数为 μ_1、而当 $R=0$ 平均数为 μ_0 的正态分布。遗憾的是，这个模型不被确认，因此若对参数没有更进一步的限制则无法估计。

形态混合模型似乎是一个思考缺失数据机制的不自然的方法。而且非常典型，我们假设数据的值（在这个例子中为 Y）是预先设定的。那么，根据数据收集程序，Y 的值可能对于我们是否确实获得想要的信息有一些影响。这样的想法与选择模型一致。另一方面，形态混合模型似乎倒转了因果关系的方向，允许缺失性影响所关注变量的分布。当然，关于因果关系方向的条件式概率是不可知的，其结果是，形态混合模型有时候比理论上更吸引人的选择模型更容易处理，特别是对于多重插补。下面作者将会给出选择模型和形态混合模型的一些例子。

第 2 节 | Heckman 的样本选择误差模型

Heckman(1976)的样本选择误差模型为缺失数据选择模型的经典例子。该模型设计用于因变量在一个线性回归模型中有一些个案有缺失、但另一些个案没有缺失的情况。一个经常使用的例子为预测女性工资的回归，而对于不在劳动力市场中的女性，其工资数据必然是缺失的。很自然地，我们会假定如果她们的工资很低的话，她们就不可能进入劳动力市场。因此，数据并不是随机缺失。

Heckman 以潜变量公式化其模型，但作者将用一个更直接的设定来处理。对于一个 n 个个案的样本 ($i=1, \cdots, n$)，令 Y 为一方差为 σ^2 的正态分布变量，且其平均数为：

$$E(Y_i) = \beta X_i, \quad [7.1]$$

其中，X_i 为自变量(包含一个截距为 1)的栏向量 M，而 β 为一个系数的列向量。目的是要估计 β。如果所有的 Y_i 都被观察到了，我们可以通过普通最小二乘回归得到 β 的 ML 估计值。然而，有些 Y_i 是缺失的。Y_i 上缺失数据的概率被假设服从 probit 模型：

$$\Pr(R_i = 0 \mid Y_i, X_i) = \Phi(\alpha_0 + \alpha_1 Y_i + \alpha_2 X_i), \quad [7.2]$$

其中 $\Phi(\cdot)$是一个标准正态变量的累进分布。除非 $\alpha_1 = 0$，否则数据不是随机缺失的，因为缺失的概率取决于 Y。

这个模型被确认(即使没有 X_i 或当 X_i 不进入该 probit 方程时)，且可以被最大似然所估计。对于一个有 Y 的观察值，其似然为：

$$\Pr(R_i = 1 \mid y_i, x_i) f(y_i \mid x_i) = [1 - \Phi(\alpha_0 + \alpha_1 y_i + \alpha_2 x_i)] \varphi\left(\frac{y_i - \beta x_i}{\sigma}\right) \sigma^{-1}, \quad [7.3]$$

其中 $\varphi(\cdot)$为一标准正态变量的密度函数。对于一个有 Y 缺失的观察值，其似然为：

$$\int_{-\infty}^{+\infty} \Pr(R_i = 0 \mid y_i, x_i) f(y \mid x_i) dy = \Phi\left(\frac{\alpha_0 + (\alpha_1 \beta + \alpha_2) x_i}{\sqrt{1 + \alpha_1^2 \sigma^2}}\right) \quad [7.4]$$

方程 7.4 服从普遍原则，即有缺失数据的观察值，其似然可以通过求所有缺失数据可能值的似然之积分而得到。使用标准数值的方法，整个样本的似然值很容易被最大化从而作出估计。

遗憾的是，这个方法所产生的估计值对于 Y 为正态分布这个假设过度敏感。如果 Y 事实上有一个偏的分布，在 Heckman 模型下得到的 ML 估计值会严重偏误，或许甚至比从一可忽略的缺失数据模型下所获得的估计值偏误更加严重(Little & Rubin, 1987)。

Heckman 也提出一个两步骤估计量，其对正态性的偏离不敏感，本质上更加容易计算，因此比 ML 更加受欢迎。然而，该两步骤方法有其自身的限制。

简而言之，两步骤如下：

(1) 估计一个缺失数据指标 R 的 probit 回归于 X 变量上；

(2) 对于 Y 有数据呈现的个案，估计一个最小二乘线性回归 Y 于 X 另加一个转换来自 probit 回归中的预测值之变量。[17]

不像 ML 方法，如果没有 X 变量，则两步骤程序不可行。此外，如果 X 变量在 proibt 回归和线性回归中都相同，参数只能被勉强地确定。为了得到合理稳定的估计值，在 probit 回归中的 X 变量必须从线性模型中排除。当然，能够令人信服地证明诸如此类的排除限制是很罕见的。甚至当所有条件都符合时，两步骤估计量在真实情况下也有可能表现得比较拙劣(Stolzenberg & Relles，1990，1997)。

在这些样本选择方法对于假设的违反有明显敏感性的前提下，我们应该如何继续做一个敏感性分析呢？对于 ML 估计量，关键在于假设为因变量 Y 的正态性。所以一个自然的方法是拟和假设不同分布的不同模型。偏的分布，如 Weibull 或 gamma 可能会是最有用的，因为对 ML 来说，正态分布的对称性最为重要。虽然在方程 7.4 中，积分可能没有一个合适的形式且可能需要数值的积分法，但是对于其他备选分布，ML 估计应该是可行的。对于两步骤估计量，关键假设是从预测的线性回归中排除某些 X 变量。一个敏感性分析可能探索对两个方程选择不同组 X 变量的结果。

第 3 节 | 形态混合模型的 ML 估计

形态混合模型很难被识别。假设我们有两个变量 X 和 Y，有四个观察到的缺失形态：

(1) X 和 Y 两者都被观察到；

(2) X 被观察到，Y 缺失；

(3) Y 被观察到，X 缺失；

(4) X 和 Y 两者都缺失。

根据这些形态中哪一种被观察到，令 $R = 1, 2, 3$ 或 4。对于这些数据的形态混合模型有一个普遍形式：

$$f(X, Y, R) = f(Y, X \mid R)\Pr(R)$$

为了使模型更加明确，我们可以假设 $\Pr(R)$ 由 p_1，p_2，p_3 和 p_4 该组值所确定。接着，我们可以假设 $f(Y, X|R)$ 为一个有着通常参数 μ_X，μ_Y，σ_X，σ_Y，σ_{XY} 的二变量正态分布。然而，我们确定这些参数中的每一个对于每一个值是不同的。问题是，当 X 被观察到而 Y 没有被观察到时，没有信息可以估计 Y 的平均数和标准差或 X 和 Y 的协方差。同样地，当 Y 被观察到而 X 没有被观察到时，没有信息可以估计 X 的平均数和标准差或 X 和 Y 的协方差。如果两个变量都缺失，我们就没有任何信息。

为了继续下去，我们必须强加一些限制在这四组参数上。令 $\theta^{(i)}$ 为形态 i 的参数组。一个简单但非常有限制的条件是，假设 $\theta^{(1)}=\theta^{(2)}=\theta^{(3)}=\theta^{(4)}$，其相当于MCAR。既然那样，形态混合模型的ML估计等同于第3章讨论过的、有可忽略的缺失数据的、正态模型的ML估计。Little(1993, 1994)提出了对其他不对应但可忽略的缺失数据产生确认的模型的限制种类。这里有一个例子。令 $\theta^{(i)}_{(Y|X)}$ 代表在 Y 为形态 i 下，Y 的条件式概率。Little称之为完全个案缺失变量限制：

$$\theta^{(2)}_{Y|X}=\theta^{(1)}_{Y|X}$$

$$\theta^{(3)}_{X|Y}=\theta^{(1)}_{X|Y}$$

$$\theta^{(4)}=\theta^{(1)}$$

对于两个存在一个变量缺失的形态，在给定该观察到的变量下，该缺失变量的条件式概率等同于完整个案形态相对应的分布。对于有两个变量缺失的形态，所有参数被假设与在完整个案形态中的那些参数相同。这个模型被确认并可以以非迭代的方法得到ML估计值。一旦所有这些参数都能获得，就可以很容易地结合它们以得到 X 和 Y 的边际分布的估计值。

第 4 节 | 形态混合模型的多重插补

形态混合模型的 ML 在目前仍是相当难以理解的。更加实际且有用得多的是组合形态混合模型与多重插补(Rubin, 1987)。最简单的策略是,首先在一个可忽略的模型下产生插补,接着以比如说一个线性来转换修正插补值。之后就可以很容易通过在线性转换中用不同的常数重复该过程来获得一个敏感性分析。

这里有一个简单的例子。再一次假设我们有两个变量 X 和 Y,但只有两种缺失数据形态:(1)完整个案,(2)Y 缺失。我们假设在每一种形态内,X 和 Y 有一个二变量正态分布。我们也相信有缺失的个案倾向于有较高的 Y 值,所以我们假设对两种形态所有参数都是相同的,除了 $\mu_Y^{(2)} = c\mu_Y^{(1)}$,其中 c 为某个大于 1 的常数。多重插补接着产生在一个可忽略的缺失数据机制下的 Y 的插补值,然后将所有插补值乘以 c。当然,要保证正常运行,我们要选一个 c 值,且该选择可以是随心所欲的。敏感性分析由对于数个不同 c 值,重插补数据及重估计模型所组成。[18]

现在,让我们将这个变成一个实际的例子。对于大专院校数据,有 98 所大专院校在因变量 GRADRAT 上有缺失数据。假设那些没有报告毕业率的大专院校,相对于那些有报

告毕业率的大专院校，有较低的毕业率，是貌似合理的。我们在第5章叙述过的、对多个插补数据的插补毕业率的平均数比没有缺失数据的大专院校的平均毕业率低了约10个百分点这个事实，支持这项假设。然而，这个差异完全来源于预测变量的差异，而且这并不构成显示数据不是随机缺失的证据。

表7.1 毕业率回归于数个变量上的不同形态混合模型

变 量	100%	90%	80%	70%	60%
CSAT	0.067	0.069	0.071	0.072	0.071
LENROLL	2.039	2.062	2.077	2.398	2.641
PRIVATE	12.716	12.542	11.908	12.675	12.522
STUFAC	−0.217	−0.142	−0.116	−0.216	−0.113
RMBRD	2.383	2.264	2.738	2.513	2.464

然而，假设在缺失及没有缺失个案间，毕业率的差异甚至会更大。毫无疑问，修正插补毕业率以使它们等于在可忽略性假设下被插补的指明的百分比。表7.1给出了插补值为原始值100%、90%、80%、70%和60%的结果。对每一个回归，会产生全新的插补。因此，横跨各栏的变化是由插补过程的随机性导致的。一般而言，系数是相当稳定的，暗示对可忽略性的偏离对结论不会有很大的影响。STUFAC系数变化最大，但在所有例子中它在统计上根本不显著。

第8章

总结与结论

在处理缺失数据的传统方法中，成列删除的问题最少。虽然成列删除可能会丢弃一大部分的数据，但除非数据不是完全随机缺失的，不然没有理由期望有误差。此外，其标准误也应该是真实标准误的适当的估计值。更进一步讲，如果你估计一个线性回归模型，对于一个自变量有缺失数据且缺失概率取决于该变量的值的情况下，成列删除是相当稳定的。如果你估计一个 logistic 回归模型，成列删除可以容忍因变量的非随机缺失或自变量的非随机缺失（但两者不能同时出现）。

相比之下，所有其他处理缺失数据的传统方法会在标准误估计值中引入误差，而且当数据为完全随机缺失时许多传统方法（如虚拟变量调整）仍会产生有偏误的参数估计值。所以成列删除是一个较为安全的方法。

如果在成列删除中必须被丢弃的数据量是无法容忍的，则有两种可备选的方法——最大似然法和多重插补。这两种方法假设数据随机缺失，这是一个较完全随机缺失更令人欣赏的较弱的假设。在颇为普遍的条件下，这些方法产生近似反映无偏误的、有效的估计值。它们也产生良好的标准误估计值和检验统计量。但不足的是，它们较大多传统方法更

难以执行，而且每次执行多元回归，它都会带来不同的结果。

如果目标是要估计一个属于由 LISREL 或相似软件包所估计的模型种类的线性模型，则最大似然可能是首选方法。当前至少有四种统计套装软件可以完成这个任务，其中最知名的为 Amos。

如果你想要估计任何类型的非线性模型，则多重插补是正确的选择，有许多不同的方法可以执行多重插补。最广泛使用的方法是假设目标模型中的变量有一多变量正态分布。通过涉及一个从数据值和参数随机抽取的重复回归的贝叶斯估计法来完成插补。当前有数个可用的套装软件可以完成这个任务。

其他做较少限制分布假设的多重插补方法目前正在发展中，但它们尚未达到理论上或计算上精细的改进水平，从而供正常化普遍使用。

也可以在数据不是随机缺失的假设下执行最大似然或多重插补，但要得到好的结果是很难处理的。这些方法对于缺失机制或有缺失数据的变量之分布的假设非常敏感。此外，没有方法检验这些假设。因此，最重要的必要条件是良好的关于缺失数据的产生机制的先验知识，且估计不可忽略的模型皆应伴随着一敏感性分析。

注释

[1] 证明很简单。我们想要估计,给定一预测变量的向量的 X 下,Y 的条件式分布 $f(Y|X)$。如果所有变量都被观察到令 $A=1$;否则,$A=0$。成列删除相当于估计 $f(Y \mid X, A=1)$。目标是要证明这个函数与 $f(Y|X)$一样。从条件式概率的定义而来,我们得到:

$$f(Y \mid X, A=1) = \frac{f(Y, X, A=1)}{f(X, A=1)}$$
$$= \frac{\Pr(A=1 \mid Y, X) f(Y \mid X) f(X)}{\Pr(A=1 \mid X) f(X)}$$

假设 $\Pr(A=1 \mid Y, X) = \Pr(A=1 \mid X)$,也就是有数据呈现于所有变量上的概率不取决于 Y,但可能取决于 X 中的任何变量。接着:

$$f(Y \mid X, A=1) = f(Y \mid X)$$

注意,这个结果可用于任何回归程序,不仅仅是用线性回归而已。

[2] 甚至当缺失数据的概率取决于 X 和 Y 两者时,有些情况下成列删除是没有问题的。令 $p(Y, X)$为一回归模型中一个或多个变量缺失数据的概率,以二分的因变量和一个自变量 X 的向量所表示的函数。如果该概率可以被因子化为 $p(Y, X) = f(Y)g(Y)$,则使用成列删除的 logistic 回归斜率为真实系数的一致的估计值(Glynn, 1985)。

[3] Glasser(1964)导出相当容易执行的方程,但是只当自变量和缺失数据形态在样本和样本间为"固定"时才有效,而这对于实际应用是一个不太可能的条件。Van Praag、Dijkstra 和 Van Velzen(1985)的方程更普遍地适用,但需要超出协方差矩阵中给定的信息:高阶动差和所有四个变量组的可得个案数目。

[4] 虽然当数据为确实缺失时虚拟变量调整方法是明显不能接受的,但它可能仍适用于未观察到的数据仅仅为不存在的个案的情况。例如,已婚受访者可能被要求评价其婚姻的质量,但该问题对于未婚的受访者就没有意义。假定我们有一个对已婚伴侣的线性方程及另一个对未婚伴侣的方程。已婚(伴侣之)方程等同于未婚(伴侣之)方程除了它具有(a)一个对应婚姻品质对因变量效应的项目以及(b)一个不同的截距。证明在这个情况下虚拟变量调整方法产生最适的估计值是容易的。

[5] 当使用条件式平均数插补时,Schafer 和 Schenker(2000)提出了一个(可)得到一致的标准误估计值(的方法)。他们宣称:在适当的条件下,他们的方法可以产生更精确的估计值,且较多重插补需要较少的运算努力。

[6] 在多变量正态假设下的最大似然,对于任何具有限四阶动差的多变量分布,产生平均数和协方差矩阵的一致的估计值(Little & Smith, 1987)。

[7] 这个变量为原始数据组中两个变量"房间支出"和"伙食变量"的总和。

[8] 对于二步骤方法,也可使用 Brown 和 Arminger(1995)叙述过的"三明治"方程得到标准误估计值。

[9] 在二变量正态性假设下估计的标准误,对于这个例子是适当的,因为数据从一个二变量正态分布抽取而来。方程为:

$$S.E.(r)=\frac{1-r^2}{\sqrt{n}}$$

虽然该样本相关系数不是正态分布的,但在这个例子中大样本数目应该确保密切近似反映正态性。因此,这些标准误可以适当地被用来建构置信区间。

[10] 对于数据扩增法,标准的不提供信息的先验分布(Schafer, 1997)为 Jeffreys 先验分布,写做 $|\sum|^{-(p+1)/2}$,其中Σ为协方差矩阵,p 为变量数目。

[11] 一个得到如此过度分散的分布的方法为使用一个自举的方法。例如,从原始数据组中有替代地取五个不同随机样本,并计算这些样本中的每一个 EM 估计值。EM 估计值接着可被用来作为五个平行系列中每一个的起始值。

[12] 一个执行这个过程的简单的方法是,将(0, 1)区间取比例于婚姻状态的每一个类别的概率的长度,分成五个次区间。对于随机数字落入的次区间指派相对应的婚姻状态。

[13] 事实上,自由度等于 L 的阶,通常但不总是为 r。

[14] 9 个地区分类如下:东部(新英格兰、中亚特兰大),566 个个案;中部(东北中央、西南中央),715 个个案;南部(南亚特兰大、东南中央、西南中央),1095 个个案;西部(山区、太平洋),616 个个案。

[15] 回归分析以 SAS 的 GLM 程序执行,使用 ABSORB 述句来处理固定效应。

[16] 产生 10 个数据组,每组 30 次反复。在插补后,插补值如有必要则被重新编码以保留原始变量可允许的值。

[17] 确切地说,额外的变量为 $\lambda(ax_i)$,其中 a 为从 probit 模型估计系数的行向量。函数$\lambda(z)$,为反 Mills 函数,被定义为 $\phi(z)/\Phi(z)$,其中 $\phi(z)$为密度函数而$\Phi(z)$为累计分布函数,这两者都用于一个标准正态变量。

[18] 在 Rubin(1987)考虑的模型中,参数的条件式先验分布被指明缺失数据形态,在给定参数于完整数据形态的条件下。这里的实例中,我仅假设缺失个案的条件式平均数为完整个案的平均数的某个倍数。此外,缺失个案的条件式方差可以被允许大于那些完整个案的(方差)以弥补较大的不确定性。

参考文献

ANGRESTI, A., and FINAY, B. (1977). "Statistical methods for the social sciences". *Engle-wood Cliffs*, NJ: Prentice-Hall.

ALLISON, P. D. (1987). "Estimation of linear models with incomplete data". In C. Clogg (Ed.), *Sociological methodology 1987* (pp. 71—103). Washington, DC: American Sociological Association.

ALLISON, P. D. (2000). "Multiple imputation for missing data". *Sociological Methods & Research*, *28*, 301—309.

BARNARD, J., and RUBIN, D. R. (1999). "Small-sample degrees of freedom with multiple imputation". *Biometrika*, *86*, 948—955.

BEALE, E. M. L., and LITTLE, R. J. A. (1975). "Missing values in multivariate analysis". *Journal of the Royal Statistical Society*, *Series B*, *37*, 129—145.

BROWNE, M. W., and ARMINGER, G. (1995). "Specification and estimation of mean and covariance structure models". In G. Arminger, C. C. Clogg, and M. E. Sobel (Eds.), *Hand Book of statistical modeling for the social and behavioral sciences* (pp. 185—249). New York: Plenum.

COHEN, J., and COHEN, P. (1985). "Applied multiple regression and correlation analysis for the behavioral sciences (2nd ed.)". *Hillsdale*, NJ: Erlbaum.

DAVIS, J. A., and SMITH, T. W. (1997). General social surveys, 1972—1996. Chicago, IL: National Opinion Research Center (producer); Ann Arbor, MI: Interuniversity Consortium for Political and Social Research (distributor).

DEMPSER, A. P., LAIRD, N. M., and RUBIN, D. B. (1977). "Maximum likelihood estimation from incomplete data via the EM algorithm". *Journal of the Royal Statistical Society*, *Series B*, *39*, 1—38.

FUCHS, C. (1982). "Maximum likelihood estimation and model selection in contingency tables with missing data". *Journal of the American Statistical Association*, *77*, 270—278.

GLASSER, M. (1964). "Linear regression analysis with missing observations among the independent variables". *Journal of the American Statistical Association*, *59*, 834—844.

GLYNN, R. (1985). *Regression estimates when nonresponse depends on the outcome variables*. Unpublished doctoral dissertation, School of Public Health, Harvard University.

GOURIEROUX, C., and MONFORT, A. (1981). "On the problem of missing data in linear models". *Review of Economic Studies*, *48*, 579—586.

GREENE, W. H. (2000). *Econometric analysis* (4th ed.), Englewood Cliffs, NJ: Prentice-Hall.

HAITOVSKY, Y. (1968). "Missing data in regression analysis". *Journal of the Royal Statistical Society, Series B*, *30*, 67—82.

HECKMAN, J. J. (1976). "The common structure of statistical models of truncated, sample selection and limited dependent variables, and a simple estimator of such models". *Annals of Economic and Social Measurement*, *5*, 475—492.

IVERSEN, G. (1985). "Bayesian statistical inference". Sage University Papers Series on Quantitative Applications in Social Science, 07—43. Thousand Oaks, CA: Sage.

JONES, M. P. (1996). "Indicator and stratification methods for missing explanatory variables in multiple linear regression". *Journal of the American Statistical Association*, *91*, 222—230.

KIM, J.-O., and CURRY, J. (1977). "The treatment of missing data in multivariate analysis". *Sociological Methods & Research*, *6*, 215—240.

KING, G., HONAKER, J., JOSEPH, A., and SCHEVE, K. (2001). "Analyzing incomplete political science data: An alternative algorithm for multiple imputation". *American Political Science Review*, *95*, 49—69. Available at http://gking.harvard.edu/stats.shtml.

LANDERMAN, L. R., LAND, K. C., and PIEPER, C. F. (1007). "An empirical evaluation of the predictive mean matching method for imputing missing values". *Sociological Methods & Research*, *26*, 3—33.

LI, K. H., MENG, X. L., RAGHUNATHAN, T. E., and RUBIN, D. B. (1991). "Significance levels from repeated p-values and multiply imputed data". *Statisica Sinica*, *1*, 65—92.

LITTLE, R. J. A. (1988). "Missing data in large surveys (with discussion)". *Journal of Business and Economic Statistics*, *6*, 287—301.

LITTLE, R. J. A. (1992). "Regression with missing X's: A review". *Journal of the American Statistical Association*, *87*, 1217—1237.

LITTLE, R. J. A. (1993). "Pattern-mixture models for multivariate incomplete data". *Journal of the American Statistical Association*, *88*, 125—134.

LITTLE, R. J. A. (1994). "A class of pattern-mixture models for normal incomplete data". *Biometrika*, *81*, 471—483.

LITTLE, R. J. A., and RUBIN, D. B. (1987). *Statisical analysis with missing data*. New York: Wiley.

LITTLE, R. J. A., and SMITH, P. J. (1987). "Editing and imputation for quantitative survey data". *Journal of the American Statistical Association*, *82*, 58—68.

MARINI, M. M. OLSEN, A. R., and RUBIN, D. (1979). *Maximum likelihood estimation in panel studies with missing data*. In K. F. Schuessler(ed.), *Sociological methodology 1980* (pp. 314—357). San Francisco: Jossey-Bass.

McCULLAGH, P. (1980). "Regression models for ordinal data(with discussion)". *Journal of the Royal Statistical Society*, *Series B*, *42*, 109—142.

McLANCHLAN, G. J., and KRISHNAN, T. (1997). *The EM algorithm and extensions*. New York: Wiley.

MOSSEY, J. J., KNOTT, K., and CRAIK, R. (1990). "Effects of persistent depressive symptoms on hip fracture recovery". *Journal of Gerontology: Medical Sciences*, *45*, M163—168.

MUTHÉN, B., KAPLAN, K., and HOLLIS, M. (1987). "On structural equation modeling with data that are not missing completely at random". *Psychometrika*, *42*, 431—462.

RAGHUNATHAN, T. E., LEPKOWSKI, J. M., VAN HOOEWYK, J., and SOLENBERGER, P. (1999). *A multivariate technique for multiply imputing missing values using a sequence of regression models*. Unpublished manuscript. Contact teraghu@umich.edu.

ROBINS, J. M., and WANG, N. (2000). "Inference for imputation estimators". *Biometrika*, *87*, 113—124.

RUBIN, D. B. (1976). "Inference and missing data". *Biometrika*, *63*, 581—592.

RUBIN, D. B. (1987). *Multiple Imputation or Nonresponse in Surveys*. New York: Wiley.

RUBIN, D. B., and SCHENKER, N. (1991). "Multiple imputation in

health-car databases: An overview and some applications". *Statistics in Medicine*, *10*, 585—598.

SCHAFER, J. L. (1997). *Analysis of incomplete Multivariate Data*. London: Chapman & Hall.

SCHAFER, J. L., and SCHENKER, N. (2000). Inference with imputed conditional means. *Journal of the American Statistical Association*, *95*, 144—154.

SCHENKER, N., and TAYLOR, J. M. G. (1996). "Partially parametric techniques for multiple imputation". *Computational Statistics and Data Analysis*, *22*, 425—446.

STOLZENBERG, R. M., and RELLES, D. A. (1990). "Theory testing in a work of constrained research design: The significance of Heckman's censored sampling bias correction for nonexperimental research". *Sociological Methods & Research*, *18*, 395—415.

STOLZENBERG, R. M., and RELLES, D. A. (1997). "Tools for intuition about sample selection bias and its correction". *American Sociological Review*, *62*, 494—507.

VACH, W. (1994). *Logistic Regression with Missing Values in the Covariates*. New York: Springer-Verlag.

VACH, W., and BLETTNER, M. (1991). "Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables". *American Journal of Epidemiology*, *134*, 895—907.

VAN BUUREN, S., and OUDSHOORN, K. (1999). "Flexible multiple mputation by MICE". Report TNO-PG 99. 054, TNO Prevention and Health, Leiden. Available at http://www. multiple-imputation. com/

VAN PRAAG, B. M. S., DIJKSTRA, T. K., and VAN VELZEN, J. (1985). "Least-squares theory based on general distributional assumptions with an application to the incomplete observations problem". *Psychometrika*, *50*, 25—36.

WANG, N., and ROBINS, J. M. (1988). "Large sample inference in parametric multiple imputation". *Biometrika*, *85*, 935—948.

WINSHIP, C., and RADBILL, L. (1994). "Sampling weights and regression analysis". *Sociological Methods & Research*, *23*, 230—257.

译名对照表

aspect	面向
autocorrelation	自相关
available case analysis	可得个案分析
Bayesian bootstrap method	贝叶斯自举法
Bayesian posterior distribution	贝叶斯后验分布
casewise deletion	个案删除
categorical independent variables	类别自变量
cell	单元格
censor	删截
chi-square	卡方
column vector	栏向量
complete-case missing-rariable restrictions	完全个来缺失变量限制
conditional distribution	条件式分布
conditional mean	条件平均数
conditional mean impution	条件式平均数插补
confidence interval	置信区间
contingency table	列联表
correlation	相关性
covariance	协方差
covariate	协变量
Cox regression	Cox 回归
Cox proportional hazards model	Cox 比例风险模型
cross-product ratio	交叉相乘比
cumulative logit model	累进的 logit 模型
cut-off point	截略点
data	数据
Data Augmentation(DA)	数据扩增法
default	默认值
denominator degrees of freedom(d. d. f)	分母自由度
departure	偏离
dependence	依赖性/相关性

dependent variable	因变量
diagnostic statistics	诊断统计量
dichotomy	二分法
dispersion	分散
disporportionate stratified sampling	非比例分层化抽样
distribution	分布
dummy variable adjustment	虚拟变量调整
effect	效应
efficient	有效的
error	误差
estimate	估计值
estimator	估计量
Expectation-Maximization(EM)	期望最大化
exploratory analysis	探索分析
factor analysis	因子分析
factor loadings	因子载荷
failure-time regression	失效时间回归
fit	拟合
fraction missing information	缺失信息比
frequency	频数
generalized linear model	广义线性模型
Heckman's selectivity bias model	Heckman 的选择性误差模型
heteroscedasticity	异方差性
higer-order moments	高阶动差
homoscedasticity	同方差性
Hot deck method	热卡方法
imputation	插补
independent variable	自变量
intended model	预设模型
iteration	迭代
joint distribution	联合分布
lags	滞后值
latent variable	潜变量

least square	最小二乘
logistic regression	logistic 回归
log-linear analysis	对数线性分析
listwise deletion	成列删除
marginal distribution	边际分布
marginal mean imputation	边际平均数插补
Markov Chain Monte Carlo(MCMC)	马尔可夫链蒙特卡尔
Maximum Likelihood(ML)	最大似然
mean substitution	平均值替换
metric	度量标准
Missing At Random(MAR)	随机缺失的
Missing Completely At Random(MCAR)	完全随机缺失的
missing data mechanism	缺失数据机制
missing data-generating mechanism	缺失数据产生机制
missing indicator method	缺失指标方法
monotonic	单调的
multinomial	多项的/多峰的
Multiple Imputation(MI)	多重插补
Multiple Imputation. by Chained Equations (MICE)	基于链式方程的多重插补
normal distribution	正态分布
null hypothesis	零假设
observed at random	随机观察的
odds	比例
overall likelihood	总体似然
overidentified model	过度识别模型
normal distribution	正态分布
p value	p 值
pairwise deletion	成对删除
panel study	面板研究
patterns	形态
poisson regression	泊松回归
population	总体

positive definite	正定的
predictive mean matching	预测均数匹配
predictor	预测(量)
probability	概率
probability density function(p. d. f.)	概率密度函数
propensity score	倾向评分
proportional odds assumption	成比例发生比假设
portion	比率
prior distribution	先验分布
ratio	比、比率
residual	残差
robust	稳健的
Root Mean Squared Error(RMSE)	均方根差
row rector	列向量
sample size	样本数
sampling variablilty	抽样变异
simulation study	模拟研究
simultaneous equations	联立方程
skewed	偏(态)的
skewness	偏度
standard deviation	标准差
standard error	标准误
stratified sampling	分层化抽样
structural equation modeling	结构方程模型
summary statistics	描述性统计
systematic error	系统性误差
t statistic	t 统计量
valid	有效的
validity	效度
variance	方差

Missing Data

上海市版权局著作权合同登记号:图字 09-2009-551